UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE INFORMÁTICA

Departamento de Sistemas Informáticos y Programación

INTEGRACIÓN DE TÉCNICAS DE CLASIFICACIÓN DE TEXTO Y MODELADO DE USUARIO PARA LA PERSONALIZACIÓN EN SERVICIOS DE NOTICIAS

MEMORIA PRESENTADA PARA OPTAR AL GRADO DE DOCTOR POR

Alberto Díaz Esteban

Bajo la dirección de los Doctores:

Pablo Gervás Gómez-Navarro
Manuel de Buenaza Rodríguez

Madrid, 2005

ISBN: 84-669-2803-0

UNIVERSIDAD COMPLUTENSE DE MADRID
DEPARTAMENTO DE SISTEMEAS INFORMÁTICOS Y PROGRAMACIÓN

INTEGRACIÓN DE TÉCNICAS DE CLASIFICACIÓN DE TEXTO Y MODELADO DE USUARIO PARA LA PERSONALIZACIÓN EN SERVICIOS DE NOTICIAS

ALBERTO DÍAZ ESTEBAN

INTEGRACIÓN DE TÉCNICAS DE CLASIFICACIÓN DE TEXTO Y MODELADO DE USUARIO PARA LA PERSONALIZACIÓN EN SERVICIOS DE NOTICIAS

Memoria que presenta para optar al grado de
Doctor en Informática

Alberto Díaz Esteban

Dirigida por los profesores

Pablo Gervás Gómez-Navarro
Manuel de Buenaga Rodríguez

Departamento de Sistemas Informáticos y Programación
Facultad de Informática
Universidad Complutense de Madrid

Abril 2005

AGRADECIMIENTOS[+]

Ha sido mucha gente la que me ha tenido que aguantar durante todos estos años en los cuales he estado trabajando en la elaboración de esta tesis. Espero acordarme de todos en estos agradecimientos.

Muchas gracias a mis directores, Pablo y Manolo, por su labor de dirección de esta tesis. Manolo me introdujo en el campo del acceso a la información y sin la participación en los diversos proyectos que él dirigió difícilmente podría haber salido adelante este trabajo. Pablo me ha apoyado y me ha ayudado mucho durante estos años, en los primeros pasos en la Universidad Europea y posteriormente contando conmigo tras su traslado a la Universidad Complutense.

Tengo que agradecer su especial apoyo a Antonio García, siempre pendiente y siempre dispuesto. Su ayuda en la parte de la evaluación cualitativa ha sido fundamental. Más allá de un colaborador, un amigo.

También ha sido muy importante la aportación de Matías Alcojor e Ignacio Acero en el desarrollo de esta tesis, su participación en los proyectos de investigación y la realización de su proyecto fin de carrera han influido mucho en el trabajo presentado aquí.

Mis compañeros de la Universidad Europea me sufrieron durante 6 años, tengo que agradecer su paciencia a Juanjo, Aliane y Oscar, compañeros de bloque de mesas, especialmente al primero por los partidos de sillón-ball para descargar adrenalina. También a Pepa, Estrella, Oscar, Bea y Miguel por la cañitas en el Miraflores. Y a los demás compañeros de macro despacho Luis, Rafa, Tomás, Alfonso, Chema, Nacho, Pedro, Javier, David y Tadu, por la compañía en múltiples cafés y comidas.

También quiero dar especialmente las gracias a los compañeros junto con los que he firmado algún artículo o he participado en algún proyecto de investigación: Inmaculada Chacón, José María Gómez, Manuel Maña, Raúl Murciano, Enrique Puertas, Beatriz San Miguel e Ignacio Giráldez.

La paciencia de mis compañeros de despacho del CES Felipe II ha sido puesta a prueba durante los últimos 3 años, sobre todo estos últimos meses. Muchas gracias a Chema y Josele por resolverme mil dudas sobre oracle, access, jsp, html, word, excel, sql, y lo que surgiera. También agradezco a Miki, Nuria, Carlos, Alfredo, Pablo, Valentín, Diego y Marivi sus comentarios y ayuda en la evaluación, aparte de pinchos de tortilla y cafés.

Tengo que agradecer al grupo de investigación GAIA de la Universidad Complutense el haberme hecho un hueco en su seno y contar conmigo para el proyecto Arcano. También por su colaboración en la evaluación de los sistemas de personalización.

La información aportada por todos los evaluadores ha sido fundamental, gracias por su ayuda anónima, aunque de tanto procesar los logs ya casi me sé el nombre de los ciento y pico.

[+] Esta investigación ha sido desarrollada parcialmente gracias a diversos proyectos de investigación del Ministerio de Industria, Iniciativa ATYCA (TS213/1999 y TS203/1999) y Programa de Fomento de la Investigación Técnica (2000/020), de la Comunidad de Madrid, Programa Regional de I+D (09/0038/99), y del Ministerio de Ciencia y Tecnología, CICYT (TIC2002-01961).

Mis amigos también han estado pendientes todos estos años, gracias por su preocupación a Carmen y Alfredo, Roberto y Amparo, Rafa y Sonia, Jorge, Manolo, Javi y Carmen, José Luis y Julia, Pepa y Paco, Carolina, Antonio y Ana, Fernando y Mª Ángeles, Gloria y Charly, Sonia y Gabriel.

El apoyo de mi familia durante todos estos años también ha sido fundamental. Gracias a mis padres Antonio y Julia, y a mis hermanos Fernando y Mario, por hacer que pudiera llegar hasta aquí. Tampoco me puedo olvidar de mis suegros Vicente y Rosario, ni de Juanvi y José, Julia y Graciela, Loli y Pati, ni de los pequeños Javier, David, Elena y Marcos.

Y por último y más importante, hubiera sido imposible conseguir terminar este trabajo sin la paciencia y el amor de Pilar y del pequeño Pablo.

A Pilar y Pablo.

RESUMEN

En los últimos años, la información disponible en formato electrónico se ha incrementado de tal manera que es muy difícil no verse saturado cuando uno intenta encontrar la información que realmente le interesa. Los contenidos Web aparecen de muy diversas maneras en distintos dominios de aplicación pero en la mayoría de ellos la forma de presentación de la información es la misma para todos los usuarios, es decir, esos contenidos son estáticos en el sentido de que no se adaptan a cada usuario desde dos puntos de vista: ni son presentados de manera diferente para cada usuario ni se adaptan a los cambios en los intereses del usuario a lo largo del tiempo. La personalización de contenidos Web trata de eliminar la sobrecarga de información mediante la adaptación de los contenidos a cada tipo de usuario y a lo largo del tiempo.

En esta tesis se muestra un enfoque integrado de personalización de contenidos Web, aplicado a servicios de noticias, basado en tres funcionalidades principales: selección de contenidos, adaptación del modelo de usuario y presentación de resultados. Todos estos procesos están basados en la representación de los intereses del usuario que estarán reflejadas en un perfil o modelo de usuario. La selección de contenidos se refiere a la elección entre todos los documentos de entrada de aquellos más interesantes para un usuario dado. La adaptación del modelo de usuario es necesaria ya que las necesidades de los usuarios cambian a lo largo del tiempo, sobre todo como resultado de su interacción con la información que reciben. La presentación de resultados consiste en, una vez seleccionados los elementos de información que más le interesan a un usuario, mostrar un documento resultado que contenga, para cada elemento seleccionado, un extracto que sea indicativo de su contenido. En particular, se ha generado un resumen personalizado por cada elemento de información seleccionado para cada usuario.

El modelo de usuario utilizado integra cuatro tipos de sistemas de referencia que permiten representar los intereses de los usuarios desde diferentes puntos de vista. Estos intereses están divididos en dos tipos: intereses a largo plazo e intereses a corto plazo. Los primeros representan intereses del usuario que permanecen constantes a lo largo del tiempo, mientras que los segundos representan los intereses que se van modificando. A su vez, el modelo a largo plazo utiliza tres métodos de clasificación que permiten al usuario definir sus necesidades de información desde 3 puntos de vista diferentes: un sistema de clasificación dependiente del dominio, donde los documentos están preclasificados por el autor del documento (p.ej.: secciones en un periódico), un sistema de clasificación independiente del dominio, obtenido a partir de las categorías del primer nivel de Yahoo! España y un conjunto de palabras clave.

Los distintos procesos de personalización se basan en técnicas estadísticas de clasificación de texto que se aplican tanto a los documentos como a los modelos de usuario. Las tareas de clasificación de texto que se utilizan están relacionadas con la recuperación de información, la categorización de textos, la realimentación y la generación de resúmenes.

La evaluación de los sistemas de personalización es especialmente compleja debido a que son necesarias las opiniones de distintos usuarios para poder obtener conclusiones relevantes sobre su funcionamiento. Para evaluar los distintos procesos de personalización se han generado varias colecciones de evaluación donde se almacenan los juicios de relevancia de varios usuarios durante varios días de utilización del sistema. Estas colecciones han permitido probar los distintos enfoques propuestos para determinar cuál de ellos era la mejor elección. Además estas colecciones pueden ser utilizadas posteriormente por otros investigadores para comparar los resultados de sus técnicas de personalización.

Las evaluaciones realizadas han mostrado que la propuesta de personalización basada en la combinación de modelos de usuario a largo y corto plazo, con resúmenes personalizados como forma de presentar los resultados finales, permite disminuir la sobrecarga de información de los usuarios, independientemente del dominio y del idioma, en un sistema de personalización de contenidos Web aplicado a servicios de noticias.

ABSTRACT

In the last years, the electronic information available has increased in such way that it is very difficult not to feel the overload when one try to find the information in which is really interested. Web content appears in many forms over different domains of application, but in most cases the form of presentation is the same for all users. The contents are static in the sense that they are not adapted to each user from two points of view: they are neither presented in a different way from each user nor capable of adapting to the interest changes of the users. Content personalization is a technique that tries to avoid information overload through the adaptation of web contents to each type of user and to the interest changes of the users.

In this thesis an integrated approach of Web content personalization applied to news services is shown. This approach is based on three main functionalities: content selection, user model adaptation and results presentation. For these functionalities to be carried out in a personalized manner, they must be based on information related to the user that must be reflected in his user profile or user model. Content selection refers to the choice of the particular subset of all available documents that will be more relevant for a given user. User model adaptation is necessary because user needs change over time, especially as result of his interaction with information. Results presentation involves generating a new result web document that contains, for each selected item, an extract that is indicative of its content. In particular, a personalized summary for each selected item for each user has been generated.

The user model integrates four types of reference systems that allow a representation of the interests of the users from different points of view. These interests are divided into two types: long term interests and short term interests. The first type represents interests of the user that remain constant over time, and the second represents the interests that are modified. The long term model uses three classification methods that allow the user to define his information needs from three different points of view: a domain dependent classification system, where the documents are pre-classified by the document author (e.g.: sections in a newspaper), an independent domain classification system, obtained of the first level categories of Yahoo! Spain, and a set of keywords.

The different personalized processes are based on statistic classification text techniques that are applied as to the documents and to the user models. The text classification tasks that are used are related with information retrieval, text categorization, relevance feedback and text summarization.

The evaluation of personalized systems is especially complex because the opinions of different users are necessary to be able to obtain relevant conclusions about system performance. To evaluate the different personalization processes some evaluation collections have been generated where the relevance judges of various users over various days are stored. These collections have made it possible to try different approaches to determine which are the best choices for this purpose. Moreover other investigators can use these collections to compare the results of their personalization techniques.

The evaluations have showed that the personalization approach based on the combination of long term and short term models, with personalized summaries as way to present the final results, achieves a certain reduction of the information overload of the users, independently of the domain and the language, in a Web content personalization system applied to news services.

ÍNDICE GENERAL

1. Introducción **12**

1.1. Motivación ... 12
1.2. Personalización de contenidos ... 13
1.3. Personalización en servicios de noticias ... 14
1.4. Objetivos ... 16
1.5. Estructura de la tesis ... 18

2. Estado del arte de la personalización de contenidos Web **20**

2.1. Introducción ... 20
2.2. Representación de los contenidos Web ... 21
2.2.1. Pre-procesamiento ... 21
2.2.2. Tipos de modelos de representación de los documentos ... 22
2.2.2.1. Modelo booleano ... 22
2.2.2.2. Modelo del espacio vectorial ... 22
2.2.2.3. Indexación de semántica latente ... 23
2.2.2.4. Modelo probabilístico ... 23
2.3. Modelado de usuario ... 24
2.3.1. Categorías propias ... 24
2.3.2. Términos ... 25
2.3.3. Categorías ... 25
2.3.4. Estereotipos ... 26
2.3.5. Redes semánticas ... 27
2.3.6. Redes neuronales ... 27
2.3.7. Largo y corto plazo ... 28
2.3.8. Filtrado colaborativo ... 29
2.4. Selección de contenidos ... 30
2.4.1. Selección basada en la fórmula del coseno del MEV ... 30
2.4.2. Selección basada en el clasificador Bayes ingenuo ... 31
2.4.3. Selección basada en el vecino más cercano ... 32
2.4.4. Selección basada en categorías ... 32
2.4.5. Selección basada en estereotipos ... 33
2.4.6. Selección basada en redes semánticas ... 34
2.4.7. Selección basada en redes bayesianas ... 34
2.4.8. Selección basada en redes neuronales ... 35
2.4.9. Selección basada en largo y corto plazo ... 35
2.5. Adaptación del modelo de usuario ... 35
2.5.1. Adaptación basada en el algoritmo de Rocchio o similares ... 36
2.5.2. Adaptación basada en el algoritmo de Bayes ingenuo o en redes bayesianas ... 37
2.5.3. Adaptación basada en el vecino más cercano ... 38
2.5.4. Adaptación basada en estereotipos ... 38

2.5.5. Adaptación basada en redes semánticas 38
2.5.6. Adaptación basada en redes neuronales 38
2.5.7. Adaptación basada en largo y corto plazo 39
2.6. Presentación de resultados 39
2.6.1. Técnicas y métodos disponibles para la presentación de resultados 39
2.6.2. Generación de resúmenes 41
2.6.3. Técnicas y métodos disponibles para la generación de resúmenes 42
2.6.3.1. Extracción de frases 42
2.6.3.2. Relaciones discursivas 44
2.6.3.3. Abstracción 45
2.7. Evaluación de sistemas de personalización 45
2.7.1. Selección y adaptación de contenidos 46
2.7.1.1. Evaluación cuantitativa 46
2.7.1.1.1. Métricas 47
2.7.1.1.2. Significancia estadística 49
2.7.1.1.3. Ejemplos de evaluación cuantitativa 49
2.7.1.2. Evaluación cualitativa o centrada en el usuario 50
2.7.1.2.1. Ejemplos de evaluación cualitativa 52
2.7.2. Presentación de resultados 53
2.7.2.1. Generación de resúmenes 54
2.7.2.1.1. Evaluación directa o intrínseca 54
2.7.2.1.2. Evaluación indirecta o extrínseca 56
2.8. Ejemplos de sistemas de personalización 58
2.8.1. SIFT 58
2.8.2. Letizia y PowerScout 58
2.8.3. Amalthea 59
2.8.4. Pefna 59
2.8.5. ANATAGONOMY 60
2.8.6. WebMate 60
2.8.7. Nakasima&Nakamura97 60
2.8.8. Balabanovic98b 61
2.8.9. PZ 61
2.8.10. News4U 62
2.8.11. METIORE 62
2.8.12. KNOME 63
2.8.13. SeAN 64
2.8.14. Browse 64
2.8.15. Shepherd *et al.* 02 65
2.8.16. IFTool 65
2.8.17. ifWeb 66
2.8.18. PIFT 67
2.8.19. NewsDude y DailyLearner 67
2.8.20. Widyantoro01 70
2.8.21. Torii 71
2.8.22. SmartGuide 71
2.8.23. Fab 72
2.8.24. P-Tango 73
2.8.25. GroupLens 73
2.9. Resumen y conclusiones del capítulo 74

3. Personalización de contenidos Web **78**

3.1. Introducción ... 78
3.2. Representación de los documentos Web ... 79
3.3. Modelado de usuario ... 79
3.3.1. Modelo a largo plazo ... 81
3.3.2. Modelo a corto plazo ... 82
3.3.2.1. Adaptación del modelo de usuario ... 82
3.3.3. Combinación de modelos a largo y corto plazo ... 85
3.4. Selección de contenidos ... 85
3.4.1. Selección con respecto al modelo a largo plazo ... 85
3.4.1.1. Selección con respecto a las categorías propias ... 85
3.4.1.2. Selección con respecto a las palabras clave ... 86
3.4.1.3. Selección con respecto a las categorías ... 86
3.4.1.4. Selección combinando los tres sistemas de referencia ... 87
3.4.2. Selección con respecto al modelo a corto plazo ... 88
3.4.3. Selección con respecto a la combinación de largo y corto plazo ... 88
3.5. Presentación de resultados ... 89
3.5.1. Resúmenes genéricos ... 91
3.5.1.1. Heurística de posición ... 91
3.5.1.2. Heurística de palabras significativas ... 92
3.5.1.3. Combinación de las heurísticas ... 92
3.5.2. Resúmenes personalizados ... 93
3.5.2.1. Heurística de personalización ... 93
3.5.2.2. Personalización utilizando las palabras clave ... 93
3.5.2.3. Personalización utilizando el modelo a corto plazo ... 94
3.5.2.4. Personalización utilizando la combinación de los modelos a corto y largo plazo ... 94
3.5.3. Combinación de las heurísticas genéricas y de personalización ... 95
3.6. Resumen de parámetros del sistema ... 95
3.7. Resumen y conclusiones del capítulo ... 96

4. Metodología de evaluación **99**

4.1. Introducción ... 99
4.2. Colección de evaluación ... 99
4.2.1. Proceso de construcción de una colección de evaluación ... 101
4.3. Tipos de evaluación ... 102
4.3.1. Evaluación cuantitativa ... 102
4.3.1.1. Significancia estadística ... 103
4.3.2. Evaluación cualitativa ... 103
4.3.2.1. Cuestionarios de evaluación ... 103
4.4. Selección de contenidos ... 105
4.4.1. Hipótesis ... 106
4.4.2. Experimentos ... 106
4.4.2.1. Experimento 1. Combinación de secciones, categorías y palabras clave, dentro del modelo a largo plazo ... 106
4.4.3. Evaluación ... 106
4.5. Adaptación del modelo de usuario ... 107
4.5.1. Hipótesis ... 107

4.5.2. Experimentos 107
4.5.2.1. Experimento 2. Combinación de modelos a corto y largo plazo. 108
4.5.2.2. Experimento 3. Combinación de largo y corto plazo, combinando secciones, categorías y palabras clave para el modelo a largo plazo. ... 109
4.5.3. Evaluación 109
4.6. Presentación de resultados 110
4.6.1. Hipótesis 111
4.6.2. Experimentos 111
4.6.2.1. Experimento 4. Generación de resúmenes personalizados. 112
4.6.2.2. Experimento 5. Combinación de heurísticas para la generación de resúmenes. 112
4.6.3. Evaluación 113
4.7. Resumen y conclusiones del capítulo 113

5. Colecciones de evaluación **116**

5.1. Introducción 116
5.2. Minicolecciones preliminares 116
5.2.1. Primera minicolección 116
5.2.2. Segunda minicolección 118
5.2.3. Tercera minicolección 118
5.3. Detalles técnicos de la construcción de las colecciones de evaluación generadas 118
5.4. Colección de evaluación 1.0 120
5.5. Colección de evaluación 2.0 124
5.6. Resumen y conclusiones del capítulo 134

6. Experimentos realizados **136**

6.1. Introducción 136
6.2. Experimentos preliminares 136
6.2.1. Primer experimento preliminar 137
6.2.2. Segundo experimento preliminar 138
6.2.3. Tercer experimento preliminar 139
6.3. Sistema de personalización de noticias 1.0 139
6.3.1. Selección de contenidos 140
6.3.1.1. Experimento 1. Combinación de secciones y palabras clave, dentro del modelo a largo plazo. 140
6.3.1.2. Resultados 140
6.3.2. Adaptación del modelo de usuario 141
6.3.2.1. Experimento 2. Combinación de modelos a corto y largo plazo. 142
6.3.2.2. Resultados 142
6.3.2.3. Experimento 3. Combinación de largo y corto plazo, combinando secciones y palabras clave para el modelo a largo plazo. 143
6.3.2.4. Resultados 143
6.3.3. Presentación de resultados 144
6.3.3.1. Experimento 4. Generación de resúmenes personalizados. 145
6.3.3.2. Resultados 145
6.3.3.3. Experimento 5. Combinación de heurísticas para la generación de resúmenes. 146
6.3.3.4. Resultados 146

6.3.4. Conclusiones del sistema de personalización 1.0 148
6.4. Sistema de personalización 2.0 149
6.4.1. Selección de contenidos 149
6.4.1.1. Experimento 1. Combinación de secciones, categorías y palabras clave, dentro del modelo a largo plazo 149
6.4.1.2. Resultados 149
6.4.2. Adaptación del modelo de usuario 151
6.4.2.1. Experimento 2. Combinación de modelos a corto y largo plazo. 151
6.4.2.2. Resultados 151
6.4.2.3. Experimento 3. Combinación de largo y corto plazo, combinando secciones, categorías y palabras clave para el modelo a largo plazo. 154
6.4.2.4. Resultados 155
6.4.3. Presentación de resultados 156
6.4.3.1. Experimento 4. Generación de resúmenes personalizados. 156
6.4.3.2. Resultados 157
6.4.3.3. Experimento 5. Combinación de heurísticas para la generación de resúmenes. 157
6.4.3.4. Resultados 158
6.4.4. Evaluación cualitativa 159
6.4.4.1. Evaluación de la interfaz 160
6.4.4.2. Valoración sobre las secciones, las categorías y las palabras clave 161
6.4.4.3. Valoración sobre la medida de la relevancia de las noticias 165
6.4.4.4. Valoración sobre los resúmenes 167
6.4.4.5. Valoración sobre la selección y la adaptación 168
6.4.4.6. Estimación global del sistema 169
6.4.4.7. Preguntas abiertas 169
6.4.5. Conclusiones del sistema de personalización 2.0 171
6.5. Resumen y conclusiones del capítulo ..... **¡Error! Marcador no definido.**

7. Discusión de resultados **175**

7.1. Introducción 175
7.2. Comparación de resultados de los distintos sistemas de personalización 175
7.3. Comparación con estado del arte 181
7.4. Extrapolación a un ámbito multilingüe 184
7.4.1. Introducción 184
7.4.2. Estado del arte 185
7.4.3. Personalización multilingüe de contenidos Web 186
7.4.4. Minicolección multilingüe 188
7.4.5. Evaluación 189
7.5. Resumen y conclusiones del capítulo 190

8. Conclusiones **192**

8.1. Principales aportaciones 192
8.2. Trabajo futuro 194

Bibliografía **195**

I. Cuestionarios de evaluación **213**

I.1. Cuestionario de evaluación utilizado en el primer experimento preliminar 213
I.2. Cuestionario de evaluación para el sistema de personalización multilingüe 217
I.3. Cuestionarios de evaluación para el sistema de personalización 2.0. 220
I.3.1. Cuestionario de evaluación inicial 220
I.3.2. Cuestionario de evaluación final 226

II. Ejemplos de noticia y modelo de usuario **238**

III. Ejemplos de resúmenes generados **240**

IV. Ejemplo de página de Yahoo! **242**

V. Esquema de la base de datos de los sistemas de personalización **243**

VI. Manual de usuario de los sistemas de personalización **245**

TABLAS

Tabla 3.1. Valores asignados a las sentencias de un documento según la heurística de posición. 92
Tabla 3.2. Parámetros del sistema con valores fijos. 95
Tabla 3.3. Parámetros ajustables del sistema. 96
Tabla 5.1. Número de secciones y palabras clave elegidas por los usuarios en la primera colección de evaluación. 121
Tabla 5.2. Pesos asignados a las distintas secciones por los usuarios. 122
Tabla 5.3. Número de noticias relevantes por usuario y por día. 123
Tabla 5.4. Número de noticias realimentadas cada día por cada usuario, de manera positiva o negativa (R+/R-). 124
Tabla 5.5. Número de noticias por día en la segunda colección de evaluación. 125
Tabla 5.6. Número de usuarios por día. 125
Tabla 5.7. Tipos de usuarios. 126
Tabla 5.8. Estadísticas sobre el número de elementos seleccionados en los perfiles de usuario. 126
Tabla 5.9. Estadísticas sobre el número de usuarios que eligieron cada uno de los métodos de selección. 127
Tabla 5.10. Pesos asignados a las distintas secciones por los usuarios. 127
Tabla 5.11. Pesos asignados a las primeras 7 categorías por los usuarios. 128
Tabla 5.12. Pesos asignados a las últimas 7 categorías por los usuarios. 128
Tabla 5.13. Número de usuarios con juicios emitidos y mensajes rebotados, por día. 129
Tabla 5.14. Número de días que unos usuarios emitieron juicios y otros no recibieron mensajes porque fueron rebotados. 130
Tabla 5.15. Estadísticas sobre el número de juicios de usuario por día y por usuario. 131
Tabla 5.16. Estadísticas sobre el número de juicios de usuario por día y por usuario, para los usuarios "válidos". 132
Tabla 5.17. Estadísticas sobre el número de usuarios "válidos" según el número de juicios por día. 133
Tabla 5.18. Estadísticas sobre el número de noticias realimentadas cada día, positiva o negativamente (R+/R-), por usuario y por día. 133
Tabla 5.19. Estadísticas sobre el número de usuarios que realimentan más noticias positiva que negativamente y viceversa, cada día. 134
Tabla 6.1. Recall y precisión normalizados para las distintas combinaciones de secciones y palabras clave, para cada día y en media. 141

Tabla 6.2. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de recall y precisión normalizados. 141
Tabla 6.3. Recall y precisión normalizados para las distintas combinaciones de largo y corto plazo, para cada día y en media. 142
Tabla 6.4. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de precisión normalizada. 143
Tabla 6.5. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de recall normalizado. 143
Tabla 6.6. Recall y precisión normalizados para las distintas combinaciones de secciones y palabras clave del modelo a largo, cuando se utiliza la combinación de largo y corto plazo, para cada día y en media. 144
Tabla 6.7. Porcentajes de mejora respecto a los valores medios de recall y precisión normalizados, para las distintas combinaciones de secciones y palabras clave del modelo a largo, cuando se utiliza la combinación de largo y corto plazo. 144
Tabla 6.8. Recall y precisión normalizados para las distintas combinaciones de los modelos a largo plazo y corto plazo para la generación de resúmenes personalizados. 145
Tabla 6.9. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de recall y precisión normalizados. 146
Tabla 6.10. Recall y precisión normalizados para los distintos tipos de resúmenes. 147
Tabla 6.11. Porcentajes de mejora de los distintos tipos de resúmenes respecto a los valores medios de recall y precisión normalizados. 147
Tabla 6.12. Recall y precisión normalizados medios para las distintas combinaciones de secciones, categorías y palabras clave dentro del modelo a largo plazo. 150
Tabla 6.13. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de recall y precisión normalizados. 150
Tabla 6.14. Precisión y recall normalizados para las distintas combinaciones de los modelos a largo plazo y corto plazo. 152
Tabla 6.15. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de precisión normalizados. 153
Tabla 6.16. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de recall normalizado. 153
Tabla 6.17. Precisión y recall normalizados para las mejores combinaciones de los modelos a largo plazo y corto plazo. 155
Tabla 6.18. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de recall y precisión normalizados. 155
Tabla 6.19. Recall y precisión normalizados para las distintas combinaciones de los modelos a largo plazo y corto plazo para la generación de resúmenes personalizados. 157
Tabla 6.20. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de recall y precisión normalizados. 157
Tabla 6.21. Recall y precisión normalizados para los distintos tipos de resúmenes. 158
Tabla 6.22. Porcentajes de mejora de los distintos tipos de resúmenes respecto a los valores medios de recall y precisión normalizados. 158
Tabla 6.23. Porcentajes de tipos de usuarios que realizaron la evaluación inicial y la evaluación final. 160

Tabla 6.24. Porcentajes de usuarios sobre cada respuesta a cada pregunta sobre la evaluación de la interfaz, en la evaluación inicial y en la evaluación final. 161

Tabla 6.25. Porcentajes de usuarios sobre cada respuesta a cada pregunta sobre la valoración de secciones, generales y palabras clave, en la evaluación inicial y en la evaluación final. 162

Tabla 6.26. Porcentajes de usuarios sobre cada respuesta a cada pregunta sobre la introducción de nuevas secciones y categorías, en la evaluación inicial y en la evaluación final. 163

Tabla 6.27. Porcentajes de usuarios sobre cada respuesta a cada pregunta sobre la selección de documentos según los distintos sistemas de referencia, en la evaluación final. 164

Tabla 6.28. Porcentajes de usuarios sobre cada respuesta a cada pregunta sobre las palabras clave, en la evaluación final. 164

Tabla 6.29. Porcentajes de usuarios sobre la preferencia en el sistema de referencia, en la evaluación final. 165

Tabla 6.30. Porcentajes de usuarios sobre la medida de la relevancia de las noticias, en la evaluación inicial y en la evaluación final. 166

Tabla 6.31. Porcentajes de usuarios sobre el interés de los usuarios por la información recibida, en la evaluación final. 167

Tabla 6.32. Porcentajes de usuarios sobre la preferencia de los usuarios en la información relacionada con la noticia utilizada para determinar la relevancia, en la evaluación final. 167

Tabla 6.33. Porcentajes de usuarios sobre cada respuesta a cada pregunta sobre los resúmenes, en la evaluación final. 168

Tabla 6.34. Porcentajes de usuarios sobre valoraciones sobre la selección y la adaptación, en la evaluación final. 168

Tabla 6.35. Porcentajes de usuarios sobre las estimaciones globales del sistema, en la evaluación final. 169

Tabla 7.1. Recall y precisión normalizados para el proceso de selección de contenidos en los sistemas de personalización 1.0 y 2.0. 177

Tabla 7.2. Comparación de la efectividad de los métodos de clasificación del modelo a largo plazo, en solitario, para la selección de contenidos en los sistemas de personalización 1.0 y 2.0. 177

Tabla 7.3. Recall y precisión normalizados para las distintas combinaciones de largo y corto plazo, para el proceso de adaptación del modelo de usuario en los sistemas de personalización 1.0 y 2.0. 178

Tabla 7.4. Recall y precisión normalizados para la combinación de secciones y palabras clave del modelo a largo plazo, junto con el corto plazo, para cada día, para el sistema de personalización 1.0. 178

Tabla 7.5. Recall y precisión normalizados para la combinación de secciones, categorías y palabras clave del modelo a largo plazo, junto con el corto plazo, para cada día, para el sistema de personalización 2.0. 179

Tabla 7.6. Recall y precisión normalizados para los distintos tipos de resúmenes generados para el proceso de presentación de resultados en los sistemas de personalización 1.0 y 2.0, con categorías y sin categorías en el proceso de selección. 180

FIGURAS

1.1. Página principal de ABC. 15
1.2. Ejemplo de mensaje enviado por ABC. 16
3.1. Modelo de usuario. 80
3.2. Ejemplo de mensaje enviado por el segundo sistema de personalización. 90
IV.1. Ejemplo de página de Yahoo! 242
VI.1. Página de inicio del sistema de personalización de noticias. 246
VI.2. Página de alta 247
VI.3. Edición del modelo de usuario (Secciones) 248
VI.4. Edición del modelo de usuario (Categorías) 249
VI.5. Edición del modelo de usuario (Palabras clave) 250
VI.6. Baja de usuario 251
VI.7. Ejemplo de mensaje. 252

Capítulo 1
INTRODUCCIÓN

1.1. Motivación

En los últimos años una gran cantidad de suministradores de contenido ha apostado claramente por la utilización de Internet como objetivo estratégico. Su éxito depende de la habilidad de los usuarios para encontrar los elementos que buscan entre miles de documentos, imágenes o productos comerciales. Estos suministradores deben ayudar a sus clientes a encontrar los productos deseados en el menor tiempo posible.

La mayoría de los servicios de acceso a la información son muy genéricos, haciendo que sea difícil para los usuarios expresar sus necesidades de información de una manera concreta. Como consecuencia, los usuarios pierden mucho tiempo buscando información relevante para sus intereses, ya que estos servicios no tienen en cuenta sus objetivos, experiencia o conocimiento. Estos factores han llevado a la aparición de servicios de información personalizada que permiten a los usuarios visualizar la información en la que están interesados. Estos servicios se suelen basar en la selección de los contenidos a mostrar en función de una estructuración de la información definida por los propios suministradores de contenidos. Por ejemplo, las secciones de un periódico o los géneros literarios de una librería permiten una clasificación de la información que el suministrador puede utilizar para mostrar al usuario sólo los documentos que pertenezcan a las secciones o géneros literarios que él seleccione.

Otra posibilidad para facilitar esta personalización en los sistemas de información es la realización de búsquedas textuales y/o el envío de la información en la que están interesados. El desafío común es encontrar una forma de relacionar los elementos de dos conjuntos: un conjunto de usuarios con unos intereses concretos y un conjunto de documentos en los que están potencialmente interesados.

La personalización de contenidos Web es un caso especial de filtrado de información que tiene como entrada un flujo dinámico de documentos y una necesidad de información. La salida del proceso de personalización es una respuesta a las necesidades del usuario, seleccionada y presentada teniendo en cuenta lo más posible sus preferencias. Los usuarios tienden a usar estos sistemas por períodos largos de tiempo, esto hace que se sientan más motivados a la hora de indicar de manera más completa y precisa la descripción de sus necesidades de información. Esta descripción es llamada habitualmente perfil o modelo de usuario y almacena varios tipos de datos que pueden ser de naturaleza heterogénea: datos personales, términos, categorías, pesos, conceptos, etc.

La personalización de contenidos ha venido utilizando distintas técnicas de selección basadas en modelos de usuario simples que almacenan generalmente los intereses de los usuarios en forma de palabras clave introducidas por el propio usuario. Estas palabras clave sirven al sistema para determinar qué documentos son los que más le interesan al usuario. Otro tipo de selección simple, habitual en sistemas de información, es el que se basa en un sistema

de categorías prefijadas por los suministradores de contenidos (p. ej.: secciones de los periódicos).

Por otro lado, una vez seleccionados los documentos más interesantes, lo habitual es mostrarle al usuario el título de cada uno de ellos, con hiperenlaces asociados, o a lo sumo, las primeras líneas. De esta forma el usuario tiene que pinchar en el hiperenlace y visualizar el documento completo para encontrar la información que realmente le interesa.

Finalmente, no todos los sistemas de personalización se adaptan al usuario a lo largo del tiempo, esto es, en algunos sistemas los intereses de los usuarios son totalmente estáticos. Este supone una limitación en la interacción de los usuarios que no permite que sus necesidades de información vayan cambiando según se va utilizando el sistema.

Por otro lado, existen en la bibliografía numerosos trabajos sobre personalización de la información, incluso existen varios trabajos recopilatorios que comparan estos sistemas desde el punto de vista de sus características y sus funcionalidades [Correia&Boavida02; van Setten01; Fink&Kobsa00; Pretschner&Gauch99; Aas97]. Sin embargo, pocos sistemas evalúan y discuten sus resultados desde el punto de vista de su efectividad. Esto es en parte debido a que es difícil determinar lo bueno que es un sistema de personalización porque incluye juicios de los usuarios y además porque cada sistema personaliza información diferente con objetivos diferentes. Además tampoco está claro cuales son las métricas más adecuadas para evaluar este tipo de sistemas.

1.2. Personalización de contenidos

En general, un sistema de personalización de contenidos está basado en 3 funcionalidades principales: selección de contenidos, adaptación del modelo de usuario y presentación de resultados [Mizarro&Tasso02; Díaz&Gervás03]. Para que estas funcionalidades se realicen de manera personalizada deben estar basadas en información relacionada con el usuario que debe estar reflejada en su perfil y debe estar disponible en el momento en el que se vaya a realizar el proceso correspondiente.

La selección de contenidos se refiere a la elección entre todos los documentos de entrada de aquellos que son más relevantes para un usuario dado, según su perfil o modelo. Para poder realizar esta selección es necesario obtener una representación de los documentos, una representación del modelo de usuario y una función que calcule la similitud entre ambas representaciones.

La adaptación del modelo de usuario es necesaria por que las necesidades de los usuarios varían con el tiempo, principalmente como efecto de la interacción con la información que reciben [Belkin97; Billsus&Pazzani00]. Por esta razón el modelo de usuario debe ser capaz de adaptarse a esos cambios de interés. Esta adaptación se realiza mediante la interacción del usuario con el sistema, a través de la cual se obtiene información para la realimentación del perfil.

La presentación de resultados consiste en, una vez seleccionados los elementos de información que le interesan a un usuario, mostrar una documento que contenga esos elementos de manera personalizada. También se deben tener en cuenta las distintas preferencias indicadas por el usuario en su perfil para generar esta información. Los resultados obtenidos se suelen plasmar en un documento Web que puede ser enviado por correo electrónico o visualizado como página Web.

Las ideas presentadas sobre personalización de contenidos se pueden generalizar a todos los tipos de información: texto, sonido, imágenes, multimedia, etc. Sin embargo, en este trabajo se va a tratar solamente con contenidos textuales por ser el formato más utilizado en la Web. Los objetivos de la personalización siguen siendo los mismos para otros formatos, pero las técnicas para procesar los contenidos cambian considerablemente. Sin embargo, muchas propuestas de personalización de estos otros formatos utilizan técnicas textuales aplicadas a los metadatos que los describen.

1.3. Personalización en servicios de noticias

Aunque las propuestas de la tesis tratan de ser los suficientemente generales como para poder ser aplicadas en distintos dominios dentro de la personalización Web, el espectro es tan amplio que ha tenido que ser acotado a un tipo de contenidos Web concreto. El dominio elegido ha sido el de las publicaciones periódicas, en particular, los servicios de noticias de los periódicos electrónicos. Este dominio tiene además unas características peculiares que van a afectar a los procesos de personalización. La principal característica es que cada día existe una colección de nuevos documentos que han de ser personalizados para todos los usuarios y que la forma de realizar la diseminación de los resultados es a través de un correo electrónico que le llega a todos los usuarios a la vez, a primera hora de la mañana.

Los servicios de noticias, y en particular los periódicos digitales, se caracterizan por la presentación de un conjunto de noticias que reflejan las noticias de cada día. Estas noticias se agrupan en secciones y cambian totalmente de un día a otro.

La web de un periódico digital se caracteriza por presentar, en su página principal, los títulos y primeras frases de los titulares del día (ver Figura 1.1). Las páginas principales asociadas a cada sección del periódico presentan la misma estructura. Se puede acceder a las noticias completas a través de hiperenlaces.

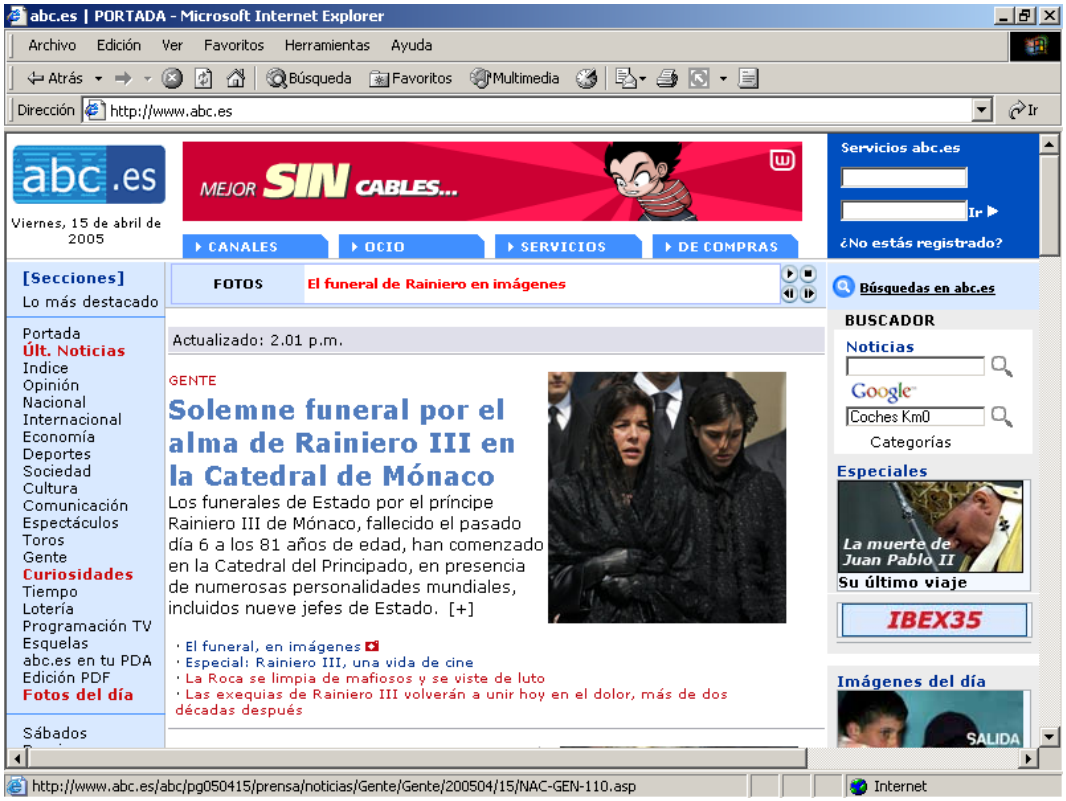


Figura 1.1. Página principal de ABC.

La mayoría de los periódicos digitales presentan sistemas de envío de noticias personalizados, de tal forma que el usuario recibe un mensaje de correo electrónico cada mañana con las noticias más interesantes según los intereses que ha reflejado a la hora de registrarse en el servicio. Estos intereses se limitan, la mayoría de las veces, a las secciones que más le interesan al usuario y en algunos casos a un conjunto de palabras clave. En cuanto a la información enviada, en la mayoría de los casos consiste en los títulos de las noticias seleccionadas (ver Figura 1.2). En algunos casos se envían además las primeras frases de las noticias. En [García *et al.* 00a] se presenta un estudio de los distintos aspectos de personalización que presentan un conjunto de servicios de noticias seleccionados.

Figura 1.2. Ejemplo de mensaje enviado por ABC.

1.4. Objetivos

En esta tesis se presenta un modelo de personalización y evaluación de contenidos Web aplicado al dominio de los servicios de noticias. En este modelo se propone la utilización de técnicas de modelado de usuario y clasificación de texto para mejorar la satisfacción final de los usuarios. Por otro lado, se plantea la evaluación desde dos puntos de vista, uno orientado al sistema que utiliza métricas de recuperación de información, y otro orientado al usuario que se basa en las respuestas de los usuarios a cuestionarios de evaluación.

Uno de los objetivos de esta tesis es mejorar la personalización de contenidos textuales en varios sentidos:

- Primero, proponer un modelo de usuario formado por varios sistemas de referencia para poder definir los intereses del usuario de una manera más completa.
- Segundo, utilizar técnicas de clasificación de texto, aplicadas a los distintos sistemas de referencia del modelo de usuario, para mejorar el proceso de selección.
- Tercero, personalizar la presentación de los resultados finales mostrados al usuario de tal forma que el esfuerzo del usuario para encontrar la información que realmente le interesa sea mínimo.
- Cuarto, permitir la adaptación de los intereses de los usuarios con el tiempo mediante la realimentación del usuario sobre la información recibida.

- Quinto, elegir técnicas que permitan cumplir los objetivos anteriores independientemente del dominio y del idioma elegido.
- Sexto, proponer una extrapolación a un sistema que maneje varios idiomas en lugar de uno.

El modelo de usuario propuesto integra cuatro tipos de sistemas de referencia que permiten representar los intereses de los usuarios desde diferentes puntos de vista. Estos intereses están divididos en dos tipos: intereses a largo plazo e intereses a corto plazo. Los primeros representan intereses del usuario que permanecen constantes a lo largo del tiempo, mientras que los segundos representan los intereses que se van modificando a lo largo del tiempo. A su vez, el modelo a largo plazo se divide en tres métodos de clasificación que permiten al usuario definir sus necesidades de información de 3 maneras diferentes: un sistema de clasificación dependiente del dominio, donde los documentos están preclasificados por el autor del documento (p.ej.: secciones en un periódico), un sistema de clasificación independiente del dominio, obtenido a partir de las categorías del primer nivel de Yahoo! España y un conjunto de palabras clave.

Los resultados se muestran a los usuarios en forma de resúmenes personalizados, de tal forma, que el esfuerzo que tiene que realizar el usuario para encontrar la información que realmente le interesa es menor que si se les mostrara sólo el título y las primeras líneas o el documento completo.

Las técnicas utilizadas para permitir la adaptación de los intereses de los usuarios se basan en la extracción de los términos más utilizados en los documentos que el usuario indica como relevantes de entre los seleccionados y presentados por el sistema.

La clasificación de texto involucra la asignación de documentos, o partes de documentos a uno o más grupos de un conjunto de ellos dado [Lewis92, Buenaga96]. Las tareas de clasificación de texto que se propone utilizar para mejorar la personalización son las siguientes: recuperación de información, categorización de textos, realimentación y generación de resúmenes.

- La recuperación de información consiste en la selección del subconjunto de documentos más adecuados a las necesidades de un usuario entre un conjunto más amplio existente en una base de datos documental [Salton&McGill83]. La necesidad del usuario suele encontrarse representada mediante una consulta formada por un conjunto de términos o palabras.
- La categorización de textos consiste en la asignación de una o más categorías preexistentes a un documento [Lewis92].
- La realimentación por relevancia se ha aplicado típicamente en entornos de recuperación de información y consiste en utilizar la opinión del usuario sobre la relevancia de los documentos recuperados para expandir la consulta inicial y repetir el proceso de recuperación [Rocchio71].
- La generación de resúmenes consiste en el proceso por el cual se identifica la información sustancial proveniente de una fuente (o varias) para producir una versión abreviada destinada a un usuario particular (o grupo de usuarios) o a una determinada tarea (o tareas) [Mani01].

Todas estas técnicas se van a aplicar al dominio de los servicios de noticias en español, pero se hará un esfuerzo porque las mismas sean extrapolables a otros idiomas y otros dominios bajo determinadas circunstancias. También se hará un primer intento de personalización multilingüe utilizando noticias en dos idiomas diferentes.

El otro objetivo de la tesis se centra en la evaluación de los sistemas de personalización de contenidos. Se propone un marco de evaluación de sistemas de personalización de contenidos que incluye unas determinadas medidas que se pueden utilizar para establecer la efectividad del sistema (evaluación cuantitativa) y la satisfacción de los usuarios (evaluación cualitativa) cuando se utilizan distintas propuestas de personalización. Se presentan varias colecciones de evaluación respecto a las cuales se evalúan las técnicas de personalización presentadas. Las colecciones de evaluación son fundamentales para poder probar la efectividad de cualquier sistema, además aportan un marco comparativo para poder establecer comparaciones entre distintas propuestas que solucionen el mismo problema. En personalización este tipo de colecciones es mucho más difícil de obtener porque cada opinión depende de cada usuario en particular.

La hipótesis de partida de esta tesis es que la propuesta de personalización basada en la combinación de modelos de usuario a largo y corto plazo, con resúmenes personalizados como forma de presentar los resultados finales, permite disminuir la sobrecarga de información de los usuarios, independientemente del dominio y del idioma, en un sistema de personalización de contenidos Web aplicado a servicios de noticias.

1.5. Estructura de la tesis

La memoria de esta tesis está organizada en ocho capítulos, incluyendo este primero de introducción, y seis apéndices.

En el Capítulo 2 se revisarán los conceptos fundamentales de los sistemas de personalización de contenidos, primero se mostrarán las distintas formas de representar contenidos, después distintas formas de representar modelos de usuario y posteriormente se dividirá el proceso de personalización en las 3 etapas mencionadas anteriormente: selección de contenidos, adaptación del modelo de usuario y presentación de resultados, y para cada una de ellas se presentarán tanto las técnicas disponibles como los métodos de evaluación utilizados para juzgar su efectividad.

En el Capítulo 3 se describirán las técnicas propuestas para realizar la personalización de contenidos Web. En primer lugar se presentará la forma elegida para representar la información manejada por el sistema. A continuación se mostrará la forma de modelar a los usuarios, fundamental en un sistema de personalización. Por último se explicarán las técnicas utilizadas en cada uno de los procesos de personalización: selección, adaptación y presentación.

El Capítulo 4 detallará la metodología de evaluación utilizada en cada uno de los procesos de personalización, indicando las métricas más adecuadas para cada tarea.

En el Capítulo 5 se describen las distintas colecciones de evaluación utilizadas en los distintos experimentos realizados a lo largo de la tesis.

En el Capítulos 6 se describen las distintas versiones desarrolladas de sistemas de personalización de noticias. En cada una de ellas se muestran las técnicas concretas utilizadas, los experimentos realizados y los resultados y conclusiones obtenidos.

En el Capítulo 7 se realiza una discusión de los resultados obtenidos, comparando los distintos sistemas de personalización entre sí y con el estado del arte. También se discute la extrapolación del sistema a un ámbito multilingüe.

Por último en el Capítulo 8 se resumen las conclusiones principales del trabajo desarrollado en esta tesis y se proponen algunas líneas de trabajo futuro.

Después de reseñar la bibliografía utilizada en el desarrollo de este trabajo, la memoria finaliza con los siguientes apéndices:

- Apéndice I: Cuestionarios de evaluación utilizados en la evaluación de las distintas versiones de sistemas de personalización.
- Apéndice II: Ejemplos de noticia y modelo de usuario utilizados.
- Apéndice III: Ejemplos de resúmenes generados.
- Apéndice IV: Ejemplo de página de Yahoo!.
- Apéndice V: Esquema de la base de datos utilizada en los distintos sistemas de personalización.
- Apéndice VI: Manuales de usuario de los distintos sistemas de personalización.

Capítulo 2
ESTADO DEL ARTE DE LA PERSONALIZACIÓN DE CONTENIDOS WEB

2.1. Introducción

En este estudio se distinguen 3 funcionalidades principales en los sistemas de personalización de contenidos Web: selección de contenidos, adaptación del modelo de usuario y presentación de resultados [Mizarro&Tasso02; Díaz&Gervás04]. Para que estas funcionalidades se realicen de manera personalizada deben estar basadas en información relacionada con el usuario que debe estar reflejada en su perfil y debe estar disponible en el momento en el que se vaya a realizar el proceso correspondiente.

Existen en la bibliografía distintas propuestas de sistemas de personalización de contenidos que emplean diversas técnicas para conseguir la adaptación de los contenidos a los usuarios. En estas técnicas existen cinco aspectos especialmente relevantes:

- La representación elegida para representar los documentos que se vayan a personalizar.
- La representación elegida para representar los modelos de los usuarios.
- La forma de seleccionar cuáles son los documentos más relevantes con respecto a un modelo de usuario.
- La forma de adaptar el modelo de usuario a través de la realimentación sobre los documentos recibidos.
- La forma de presentar los resultados obtenidos en el proceso de selección.

El resto del capítulo se distribuye de la siguiente manera: en el apartado 2.2 se van a describir las técnicas que se utilizan para representar la información manejada por los sistemas de personalización de contenidos, en el apartado 2.3 las técnicas que sirven para modelar a los usuarios, en el apartado 2.4 se mostrarán las técnicas habituales utilizadas en el proceso de selección de contenidos, en el apartado 2.5 las utilizadas en el proceso de adaptación del modelo de usuario, y en el apartado 2.6 se mostrarán las técnicas correspondientes a la presentación de resultados.

En el apartado 2.7 se mostrarán los distintos métodos de evaluación disponibles para juzgar los distintos procesos de personalización de contenidos.

En el apartado 2.8 se mostrarán los detalles más relevantes de los distintos sistemas analizados.

Por último, en el apartado 2.8 se mostrará un resumen y las conclusiones del capítulo.

2.2. Representación de los contenidos Web

La obtención de una representación del contenido textual de los documentos Web consiste en asignar una serie de descriptores de contenido a cada uno de los documentos. Estos descriptores suelen ser habitualmente términos, donde usaremos término en lugar de palabra puesto que estos descriptores de contenido pueden ser partes de palabras, palabras o combinaciones de palabras. El conjunto de descriptores asociado a un documento representa, por tanto, el contenido de un documento.

A este proceso de obtener la representación de los documentos se le suele llamar indexación de documentos, puesto que lo que se obtiene es una serie de índices (descriptores) que representan cada uno de los documentos.

Los descriptores pueden ser también metadatos, que describan el contenido del documento en lugar de formar parte de él. Por ejemplo: un video digitalizado es el contenido, mientras que la descripción en forma textual de su contenido es un metadato asociado al mismo.

Existen distintas técnicas para asociar términos a los contenidos de los documentos que se van a describir a continuación [Salton&McGill83; Aas97]. En primer lugar, se suele realizar un pre-procesamiento de los documentos para obtener los términos que aparecen en los mismos. Este aspecto se tratará en el apartado 2.2.1. En segundo lugar, apartado 2.2.2, se tratarán distintos modelos de representación de los documentos.

2.2.1. Pre-procesamiento

La idea de utilizar documentos de la Web como fuentes de información choca con la heterogeneidad extrema de los documentos HTML [Martínez *et al.* 01]: no comparten el mismo formato, ni aún procediendo de la misma fuente de información. Esto requiere un esfuerzo extra para preprocesar estos documentos y obtener la colección inicial de documentos "limpia", es decir, sin etiquetas y sólo con la información que nos interesa.

Por tanto, el primer paso consiste en pre-procesar el documento Web y extraer del mismo la información que va a ser utilizada para representarlo. Esta información generalmente será simplemente el texto que describe el documento, aunque en algunos casos se pueden obtener metadatos que aporten información adicional sobre el contenido Web.

Una vez obtenido el texto asociado al documento se procede a la obtención de las palabras que aparecen en el mismo. De estas palabras son eliminadas todas aquellas que aparezcan en una lista de palabras "vacías" puesto que estas palabras no aportan nada significativo al contenido de los documentos (pronombres, preposiciones, conjunciones, etc.) [Salton&McGill83].

Por último, se realiza un proceso de eliminación de sufijos y prefijos de las palabras, para obtener los términos que representan raíces de palabras. El algoritmo más habitual es el de Porter (inglés) [Porter80]. Existe una adaptación del algoritmo de Porter al español [Acero *et al.* 01].

En algunos trabajos también se tratan como descriptores combinaciones de términos consecutivos (n-gramas) [Sorensen&Elligott95].

2.2.2. Tipos de modelos de representación de los documentos

La mayoría de los modelos utilizados para la representación de documentos en personalización de contenidos son los utilizados habitualmente en recuperación de información. Estos modelos también se utilizan en muchas ocasiones para representar los modelos de los usuarios.

2.2.2.1. Modelo booleano

El modelo de representación más básico es el modelo booleano [Salton&McGill83] donde cada término es tratado como una variable booleana, de tal forma que tiene el valor VERDADERO si la palabra aparece en el documento y FALSO en caso contrario. Los términos no tienen peso asociado por lo que todos los términos distintos que aparecen en un documento son considerados igualmente importantes a la hora de representar el contenido del mismo.

2.2.2.2. Modelo del espacio vectorial

El método más habitual utilizado para representar los documentos es el modelo del espacio vectorial, a partir de ahora MEV [Salton&McGill83]. En este modelo los documentos son representados como vectores de pesos de términos. Si se utilizan m términos distintos para la representación, entonces un documento D_d se representa con un vector de m dimensiones $D_d = (w_{d1}, ..., w_{dm})$, donde cada w_{dj} representa el peso del término j en el documento d. La dimensión del espacio vectorial viene dado por el número de términos distintos que aparezcan en los documentos.

Los pesos de los términos representan la importancia del término como representante del contenido del documento. Existen varias propuestas para el cálculo de dichos pesos donde la mayoría de ellas combinan dos aspectos: la frecuencia de apariciones de un término dentro de un documento (tf, *term frequency*) y la frecuencia de documentos en los que aparece un término (df, *document frequency*) o su inversa (idf, *inverse document frequency*). La más habitual es la combinación tf · idf. La justificación de esta medida está en el hecho de que un término es más representativo del contenido de un documento cuantas más veces aparezca en el mismo y cuantos menos documentos distintos lo contengan.

En realidad la combinación tf · idf es más una familia de tipos de pesos que un peso en concreto, ya que estos valores pueden ser calculados de distinta forma [Salton&Buckley88]. En todo caso, la manera más habitual de calcularlos es con tf como la frecuencia de apariciones de un término dentro de un documento e $idf = \log_2(N/df)$, siendo N el número de documentos de la colección.

Por otro lado, la similitud entre una consulta c_k y un documento d_j se calcula mediante la fórmula del coseno entre dos vectores de pesos de términos (ecuación (2.1)) [Salton&McGill83]:

$$sim(d_j, c_k) = \frac{\sum_{i=1}^{m} wd_{ji} \cdot wc_{ki}}{\sqrt{\sum_{i=1}^{m} wd_{ji}^{\,2} \cdot \sum_{i=1}^{m} wc_{ki}^{\,2}}} \qquad (2.1)$$

El problema de la utilización de términos individuales es la ambigüedad: un mismo término puede tener significados diferentes que no se pueden distinguir con este tipo de representación. Además también puede ocurrir que tengamos varios términos que representen el mismo significado. La utilización de frases o grupos de términos puede solucionar este problema, sin embargo, su obtención es bastante compleja y no asegura una mejora en la efectividad del modelo.

Además este modelo tiene la desventaja de que considera los términos como mutuamente independientes. Sin embargo, considerar la dependencia entre términos puede ser una desventaja ya que la mayor parte de estas dependencias son locales al contexto y considerarlas como generales podría llevar a peores resultados [Baeza-Yates&Ribeiro-Neto99].

2.2.2.3. Indexación de semántica latente

La indexación de semántica latente (LSI, *Latent Semantic Indexing*) [Dumais *et al.* 88; Deerwester et al 90; Dumasi04] es una extensión del modelo del espacio vectorial que fue diseñado para incorporar información semántica. Se basa en la utilización contextual de las palabras en un *corpus* de gran tamaño para extraer y representar el significado de las palabras y conjuntos de palabras, haciendo uso de cálculos estadísticos. Surgió como una herramienta para la indexación y recuperación de información que resolvía el problema de la polisemia.

LSI usa descomposición en valores singulares (DVS), una técnica muy relacionada con la descomposición en autovectores y el análisis factorial. DVS descompone una matriz de tamaño mxn (documentos x términos) en un producto de matrices, donde una de las matrices (M) es una matriz diagonal con k valores singulares (autovalores) distintos (k=min{m,n}) y las otras dos son matrices formadas por vectores singulares (autovectores) [Golub&Loan96]. Los autovalores son únicos y el número de ellos distinto de cero es el rango r de la matriz M (r <= k).

Posteriormente se puede aproximar la matriz M por otra M_p de rango p <= r, calculada para los p primeros autovalores. Esto permite reducir la dimensión del espacio vectorial a un espacio de dimensión p, donde se captura mejor la estructura que asocia términos y documentos.

2.2.2.4. Modelo probabilístico

Uno de los modelos fundamentales y más representativos de los modelos probabilísticos en recuperación de información es el *Binary Independence Retrieval Model* (BIR) [Robertson&Sparck Jones76; Baeza-Yates&Ribeiro-Neto99]. En este modelo los documentos se representan con una indexación booleana de términos. El sistema calcula la similitud entre un documento y una consulta en función de las probabilidades de que el documento d pertenezca al conjunto de documentos relevantes a la consulta c y de que no pertenezca (ecuación (2.2)).

$$\mathrm{si}\, m(d,c) = \frac{P(R \mid d)}{P(\overline{R} \mid d)} \qquad (2.2)$$

Esta similitud, a través del teorema de Bayes, se convierte en la ecuación (2.3), donde P(d|R) indica la probabilidad de seleccionar el documento d del conjunto R de documentos relevantes y P(R) indica la probabilidad de seleccionar un documento relevante de entre los documentos de la colección. Además se elimina el último factor porque es constante para todos los documentos.

$$P(x \mid y) = \frac{P(y \mid x) \cdot P(x)}{P(y)} \Rightarrow sim(d,c) = \frac{P(d \mid R) \cdot P(R)}{P(d \mid \overline{R}) \cdot P(\overline{R})} \approx \frac{P(d \mid R)}{P(d \mid \overline{R})} \quad (2.3)$$

Si además se hace la suposición de que los términos de los documentos son condicionalmente independientes, el resultado final se convierte en productos de probabilidades de que los términos del documento aparezcan en documentos relevantes y no relevantes (ecuación (2.4)).

$$P(d \mid R) = \prod_{i=1} P(t_i \mid R) \Rightarrow sim(d,c) = \prod_{i=1} \frac{P(t_i \mid R)}{P(t_i \mid \overline{R})} \quad (2.4)$$

Finalmente, se suele utilizar una transformación logarítmica para escalar los resultados (ecuación (2.5)).

$$sim(d,c) = \log \frac{P(d \mid R)}{P(d \mid \overline{R})} \Rightarrow sim(d,c) = \sum_{i=1} \log \frac{P(t_i \mid R)}{P(t_i \mid \overline{R})} \quad (2.5)$$

Estas probabilidades se pueden calcular mediante entrenamiento previo o se pueden estimar sus valores iniciales cuando no hay documentos relevantes. En todo caso se pueden ir ajustando a través de la realimentación del usuario.

Este modelo básico se puede mejorar mediante la utilización de los valores de las frecuencias de los términos y de la longitud de los documentos. En Okapi [Sparck Jones *et al.* 98] se modelan las frecuencias de los términos dentro de los documentos como una mezcla de dos distribuciones de Poisson, una para documentos relevantes y otra para no relevantes.

2.3. Modelado de usuario

El modelo o perfil de usuario es el que almacena los distintos intereses que tiene cada usuario. Esta información es la que se utiliza para seleccionar qué documentos son relevantes para cada usuario.

Existen en la bibliografía distintas formas de definir los intereses de los usuarios que dependen básicamente de la cantidad y el tipo de información que se utilice para la representación de dichos intereses. Existen estudios [Aas97; Pretschner&Gauch99; Van Setten01; Bueno02; Cruz *et al.* 03] que recogen distintas técnicas aplicadas y sistemas que las utilizan. A continuación se muestran las técnicas consideradas más importantes, así como algunos sistemas que las implementan. En el apartado 2.8 aparecen todos los sistemas analizados en este trabajo, junto con sus características más interesantes.

2.3.1. Categorías propias

La utilización de categorías propias o predefinidas es la técnica más sencilla de personalización, aunque es dependiente del dominio de aplicación. Los documentos son pre-asignados por expertos en el dominio del sistema de categorías. Por ejemplo: las secciones de un periódico serían un sistema de categorías predefinidas.

En el perfil de usuario se almacenan las categorías en las que está interesado el usuario. Estas categorías también pueden tener un peso asignado por el propio usuario.

My Yahoo! (http://my.yahoo.com) es un ejemplo de sistema de este tipo, donde el usuario puede seleccionar las categorías de Yahoo! en las que está interesado.

En otros trabajos se utiliza esta técnica combinada con otras que representan los intereses de los usuarios desde otros puntos de vista. Por ejemplo, en PZ [Veltmann98], New4U [Jones *et al.* 00] y P-Tango [Claypool *et al.* 99] se utilizan las secciones de los periódicos, combinadas con otras fuentes de información, para personalizar periódicos digitales (ver apartado 2.8 para más detalles).

2.3.2. Términos

El método más simple y más habitual de representar al usuario, cuando se trata de establecer sus preferencias con respecto al contenido de los documentos, consiste en que el usuario introduzca un conjunto de palabras clave o términos que representen sus intereses. Adicionalmente se le puede asignar un peso a cada una de esas palabras para determinar con más exactitud cuál es la importancia asociada a cada una de ellas. Estos conjuntos de palabras clave con pesos suelen ser tratados como vectores en el MEV.

Un ejemplo de sistema de este tipo es SIFT [Yan&García-Molina95]. En este sistema de filtrado, utilizado con grupos de noticias, cada usuario puede construir varios perfiles, donde cada perfil puede ser expresado en uno de los siguiente modelos de recuperación de información: modelo booleano o MEV.

Otros ejemplos de sistemas que utilizan esta técnica son: Letizia [Lieberman95], PowerScout [Lieberman *et al.* 01], Amalthea [Moukas96], Pefna [Kilander *et al.* 97], ANATAGONOMY [Sakagami&Kamba97], [Nakasima&Nakamura97], WebMate [Chen&Sycara98], PZ [Veltmann98], [Balavanovic98b], News4U [Jones *et al.* 00], METIORE [Bueno01, 02] (ver apartado 2.8 para más detalles).

2.3.3. Categorías

Otra forma de definir los intereses de los usuarios es a través de categorías representadas por conjuntos de términos obtenidos de conjuntos de documentos clasificados previamente en cada una de esas categorías. Esta forma de definir los intereses es similar a tener varios perfiles compuestos por vectores de pesos de términos, que representan distintos temas de interés del usuario. Sin embargo, la diferencia estriba en que los perfiles pueden cambiar mediante la realimentación del usuario, mientras que la representación de las categorías es, en principio, estática. Los usuarios eligen las categorías que mejor representan sus intereses y les pueden asignar un peso para determinar más exactamente la importancia que les otorgan.

La determinación de similitudes entre documentos y categorías es estudiada por un tipo de técnica de clasificación de texto denominada categorización de texto [Sebastiani99; Yang99]. Esta técnica consiste en determinar la asignación de documentos a un conjunto de categorías previamente definidas. Los documentos pueden ser de cualquier tipo de elemento de información con contenido textual: noticias, artículos, páginas web, correos electrónicos, etc. Las categorías pueden tomarse a partir de cualquier sistema de categorías previamente definido: categorías de una biblioteca, categorías de directorios de Internet, etc. Los documentos pueden pertenecer a varias categorías.

Las primeras propuestas de sistemas de categorización estaban basadas en sistemas expertos, que permitían clasificar los documentos en base a una serie de reglas codificadas manualmente. Sin embargo, los sistemas de categorización basados en técnicas de aprendizaje máquina tienen mayor importancia en la actualidad debido a la mayor cantidad de información que se maneja. Con este paradigma, un proceso inductivo construye automáticamente

un clasificador por "aprendizaje". Este proceso parte de una serie de documentos clasificados manualmente (colección de entrenamiento) para extraer las características que representan a cada una de las categorías que maneja el sistema. Las principales ventajas de este enfoque son su efectividad, un considerable ahorro de esfuerzo humano y la independencia del dominio de aplicación. Existe una gran variedad de técnicas de aprendizaje aplicadas a la categorización de texto, la mayoría de ellas se pueden encontrar descritas en algunos trabajos recopilatorios [Sebastiani99; Yang99; Sebastiani02].

Existen multitud de sistemas que aplican técnicas de categorización de texto para clasificar documentos dentro de sistemas de categorías [Sebastiani02]. En particular existen varias propuestas [Mladenic98a, 98b; Labrou&Finin00; Gómez01] que utilizan la jerarquía de categorías de Yahoo! para construir un clasificador, utilizando las páginas indexadas en dicho sistemas de categorías como los elementos para entrenar el clasificador. En particular los resultados en [Labrou&Finin00] sugieren que las descripciones breves de las entradas que aparecen en las páginas asociadas a las categorías de Yahoo! son las que producen mejores resultados a la hora de representar las categorías. Estas descripciones son producidas por los suministradores o indexadores humanos de la página.

Por otro lado, en los sistemas de personalización es muy común la utilización de temas de interés en forma de grupos de términos introducidos por los usuarios, pero la utilización de categorías es mucho más rara. De hecho, en los sistemas revisados no se ha encontrado ninguno que utilice categorías para representar modelos de usuario.

2.3.4. Estereotipos

Los estereotipos están asociados a tipos de usuarios, donde cada tipo incluye a un grupo de usuarios que comparten intereses según unos criterios determinados [Rich79; Rich83]. Los estereotipos asumen el principio de que si un usuario pertenece a un determinado grupo, entonces tendrá características y/o comportamientos semejantes a los miembros de ese grupo, bajo un determinado conjunto de circunstancias.

Los principales componentes de un estereotipo son: las características comunes a todos los usuarios que pertenecen al mismo grupo, las condiciones de activación (*triggers*), que hacen que a un usuario se le aplique un determinado estereotipo y la base de conocimiento de reglas que determinan las probabilidades asociadas a las predicciones de interés de los usuarios por cada tipo de documento.

Razonar sobre estereotipos consiste en evaluar las reglas de activación cada cierto tiempo, y si se cumple alguna condición para el usuario actual, entonces se aplican las reglas correspondientes al estereotipo en el perfil de ese usuario para determinar los documentos que pueden interesar al usuario. Actualmente los estereotipos están casi siempre basados en observaciones empíricas sobre los usuarios (análisis de los datos del usuario (entrevistas), datos de uso del sistema, etc.).

La efectividad de este enfoque depende de la calidad de los estereotipos: número de ellos diferentes, acierto en la distribución de los usuarios, calidad de las reglas que determinan las probabilidades de interés asociadas a cada estereotipo, etc. También es fundamental que la información disponible sobre la población de los usuarios y su división en tipos de usuarios sea de buena calidad.

En KNOME[Chin89] se presenta un sistema de ayuda para el sistema operativo UNIX. La ayuda que el sistema ofrece al usuario se basa en el conocimiento que tiene el sistema del usuario. En función de las preguntas que realiza el usuario, el sistema es capaz de inferir el

conocimiento de los usuarios y responder de una manera más adecuada. Se maneja un sistema de doble-estereotipo donde un conjunto de estereotipos representa el nivel de experiencia de los usuarios (novato, principiante, intermedio y experto) y otro representa el nivel de dificultad de la información (simple, usual, complejo y avanzado). KNOME almacena información, en forma de proposiciones, sobre lo que los usuarios saben y no saben. Dependiendo del diálogo con el sistema, los modelos de los usuarios van añadiendo hechos a su perfil a través de inferencias codificadas en forma de reglas.

Otros ejemplos de sistemas que utilizan esta técnica son: el trabajo presentado en [Benaki *et al.* 97], y el sistema SeAN [Ardissono *et al.* 99a; 99b; 01] (ver apartado 2.8 para más detalles).

2.3.5. Redes semánticas

Una red semántica es un sistema de organización del conocimiento en el que los conceptos se estructuran en forma de red, donde los nodos son los conceptos y los enlaces representan relaciones entre los conceptos. Estas relaciones pueden ser de muchos tipos: concepto más amplio, concepto más específico, sinónimo, asociado o relacionado, todo-parte, causa-efecto, parte-de, etc. Las redes semánticas son grafos orientados que proporcionan una representación declarativa de objetos, propiedades y relaciones [Quillian68].

En IFTool [Asnicar *et al.* 97] se describe un sistema de filtrado basado en un modelo de usuario que describe los "intereses" y "no intereses" de un usuario. Más concretamente, está formado por dos redes semánticas, una que representa los "intereses" del usuario y otra que representa los "no intereses", es decir, información sobre la que no está interesado el usuario. Cada red semántica contiene nodos que se corresponden con términos (conceptos) encontrados en los documentos y donde los arcos unen términos que co-ocurren en el mismo documento. Cada nodo tiene un peso que puede ser positivo o negativo dependiendo de si se ha extraído de un documento juzgado como interesante o como no interesante. Cada arco está caracterizado con un peso que indica la frecuencia de la co-ocurrencia de los términos en los documentos previamente analizados. La utilización de la co-ocurrencia disminuye el problema de la polisemia de las palabras ya que este tipo de relación permite asociar a cada término un "contexto pragmático" que ayuda a la desambiguación de un término.

Otros ejemplos de sistemas que utilizan esta técnica son: ifWeb [Asnicar&Tasso97] y PIFT [Asnicar *et al.* 97] (ver apartado 2.8 para más detalles).

2.3.6. Redes neuronales

Las redes neuronales artificiales (ANN) son sistemas paralelos para el procesamiento de la información, inspirados en el modo en el que las redes de neuronas biológicas del cerebro procesan información. Se caracterizan por un conjunto de nodos de entrada que representan las posibles entradas del problema a resolver, un conjunto de nodos de salida que representan estados de salida del problema y un algoritmo de aprendizaje que sirve para actualizar el estado del sistema durante el entrenamiento [Mitchell97].

En el contexto del modelado de usuario, cada usuario puede tener un perfil caracterizado por un conjunto de valores asociados a atributos. La red va actualizando los pesos de sus enlaces, los cuales representan el estado de la red, según va avanzando el proceso de entrenamiento. Este conjunto de pesos representa el perfil del usuario en cada momento. Una red neuronal bien entrenada puede reconocer patrones de entrada incompletos, incluso si no se

dispone de todo el conocimiento sobre los atributos asociados a cada usuario [Shepherd *et al.* 02].

Browse [Jennings&Higuchi92] es un lector de grupos de noticias que utiliza redes neuronales para representar los intereses del usuario. La red neuronal almacena asociaciones entre parejas de palabras (no necesariamente adyacentes en el texto) de tal manera que llega a ser sensible a probabilidades de co-ocurrencia. Cada nodo está asociado a una palabra que aparece en artículos que ha leído el usuario y tiene una energía que determina lo importante que es esa palabra para el usuario. Por otro lado, los enlaces representan el grado de asociación entre palabras, en el sentido de co-ocurrencia y también tienen un peso que indica la importancia para el usuario. Las primeras 300 palabras de los artículos (filtradas de palabras vacías) representan los atributos del documento y son las que sirven para construir/ajustar la red.

Otros ejemplos de sistema que utiliza esta técnica es el presentado en [Shepherd *et al.* 02] (ver apartado 2.8 para más detalles).

2.3.7. Largo y corto plazo

En otro tipo de sistemas, se plantean modelos de usuario donde los usuarios tienen preferencias a distintos niveles: persistente (o a largo plazo), esto es, basado en una necesidad de información que permanece más o menos estática a lo largo del tiempo, y efímera (o a corto plazo), en este caso, la necesidad es algo más puntual y no permanece estática a lo largo del tiempo. Dependiendo de como de modificable se considere el largo plazo éste puede ser estático y no variar a lo largo del tiempo, o ser dinámico, y variar a lo largo del tiempo. En todo caso la variación en el largo plazo es mucho más lenta que la que se produce en el corto plazo.

La utilidad de la inducción de dos modelos de usuario separados fue estudiada en [Chiu&Webb98] en el contexto del modelado de estudiantes. Las técnicas y el dominio son diferentes a las de esta tesis doctoral pero la motivación subyacente del modelo de usuario dual es similar. En general, el modelado de usuario es una tarea con características inherentemente temporales. Se puede asumir que los datos más recientes recogidos de la interacción con el usuario reflejan mejor el conocimiento, habilidades o preferencias de un usuario que aquellos basados en interacciones producidas en momentos anteriores. Sin embargo, restringir los modelos a datos recientes puede llevar a modelos que clasifiquen instancias que son similares a datos recientes con gran precisión pero den malos resultados con instancias que se desvíen de los datos utilizados para inducir el modelo inicial. Para evitar este problema, Chiu y Webb usan un modelo dual que clasifica las instancias consultando primero en un modelo a corto plazo entrenado con datos recientes y si éste no es capaz de realizar la clasificación, se utiliza un modelo a largo plazo dinámico basado en una interacción durante mucho más tiempo.

Un ejemplo de aplicación de esta idea es NewsDude [Billsus&Pazzani99a, 99b]. Este sistema construye un programa personalizado de noticias basado en un sintetizador de voz (en teoría para una radio de un coche). Además de representar por separado los intereses a corto y largo plazo, ambos de manera dinámica, tiene en cuenta las noticias que ya han sido presentadas al usuario para evitar presentar la misma información dos veces.

La representación de los intereses del usuario se basa en técnicas de aprendizaje máquina [Webb *et al.* 01]. En particular, se utilizan las 100 últimas noticias sobre las que el usuario ha emitido algún tipo de juicio para representar los intereses a corto plazo del usuario. Estas noticias se representan como vectores de pesos de términos y tienen asociada una puntua-

ción, entre 0 y 1, que se obtiene de la valoración de los usuarios de la siguiente forma: si el usuario elige "no interesante" => 0.3 * pl, si elige "interesante" => 0.7 + 0.3 * pl, si pregunta por más información => 1.0 (siendo pl la proporción del artículo que el usuario ha oído).

Por otro lado los intereses a largo plazo sirven para modelar las preferencias generales del usuario y se utilizan cuando un nuevo artículo no puede ser clasificado por el modelo a corto plazo. En este caso, se representan las noticias como vectores de booleanos, donde cada posición del vector está asociada a una determinada palabra (característica). Estos términos se eligen manualmente como buenos indicadores de las noticias que aparecen más comúnmente. Se seleccionaron aproximadamente 200 palabras pertenecientes a distintos temas como atributos para un clasificador bayesiano ingenuo. También se almacena la opinión del usuario ("interesante" o "no interesante") sobre las noticias.

Otros ejemplos de sistemas que utilizan esta técnica son: DailyLearner [Billsus&Pazzani00], [Widyantoro01], Torii [Mizarro&Tasso02] y SmartGuide [Gates *et al.* 98] (ver apartado 2.8 para más detalles).

2.3.8. Filtrado colaborativo

La idea en la que se basa esta técnica es la siguiente: los usuarios indican sus opiniones suministrando valores de relevancia a varios contenidos de información, y el filtro colaborativo correlaciona estos valores con los suministrados por otros usuarios, para realizar futuras predicciones. Si un usuario da un valor alto de relevancia a un nuevo documento y hay varios usuarios que han realizado anteriormente juicios similares, entonces el sistema recomendará este nuevo documento a estos usuarios. Cuando se utiliza filtrado colaborativo el perfil de usuario almacena los valores de relevancia asignados por el usuario a cada uno de los elementos de información que ha valorado. Además, en el filtrado colaborativo estos valores son compartidos con otros usuarios para que los puedan usar para sus propias predicciones.

Los primeros sistemas requerían que el usuario explícitamente efectuará las valoraciones (P.ej.: Tapestry [Goldberg *et al.* 92]). Posteriormente surgieron sistemas donde el esfuerzo del usuario disminuía considerablemente (P.ej.: GroupLens[Resnick *et al.* 94; Konstan *et al.* 97]). E incluso han surgido sistemas donde las valoraciones se extraen de manera implícita a partir de las acciones del usuario [Terveen *et al.* 97].

La principal ventaja de este tipo de sistemas es que no considera el contenido de los documentos sino que únicamente se basa en las valoraciones de los usuarios. Sin embargo, la utilización en solitario de un sistema de filtrado colaborativo puede ser poco efectiva debido a varias razones: el problema del primer evaluador, el problema de la escasez de valoraciones y el problema de la oveja negra [Claypool *et al.* 99].

El primer problema se produce cuando aparece por primera vez un contenido de información, ya que no existen valoraciones de los usuarios en las que basarse para realizar predicciones. El segundo es consecuencia de que normalmente existen muchos más elementos que valorar que usuarios que los valoren, lo cual puede producir que no haya suficiente número de valoraciones sobre un elemento de información que permita realizar predicciones adecuadas. El tercer problema aparece con usuarios que no poseen un comportamiento consistente con ningún grupo de usuarios. Estos usuarios raramente reciben predicciones adecuadas a sus intereses.

Hay experimentos que han mostrado que se pueden mejorar este tipo de sistemas utilizando filtros basados en contenido [Balabanovic&Shoham97; Claypool *et al.* 99; Good *et al.* 99]. Mediante la combinación de filtros colaborativos y basados en contenido se pueden

aprovechar los beneficios de los filtros basados en contenido, los cuales incluyen predicciones iniciales que afectan a todos los elementos de información y a todos los usuarios, y además aprovechar los beneficios de las predicciones del filtrado colaborativo cuando el número de usuarios y de evaluaciones se incrementa.

Fab [Balabanovic97&Shoham97] es un sistema de recomendación de páginas Web, basado en filtrado colaborativo, donde los documentos se representan mediante los 100 términos con mayor peso $tf \cdot idf$. Cuando un usuario nuevo llega al sistema se le ofrece una serie de páginas aleatorias de entre un conjunto de páginas que son las que más le interesan al resto de usuarios que utilizan el sistema. De esta forma el usuario no empieza con un perfil vacío. Para ello se almacena la media de las evaluaciones de los usuarios en un perfil global.

Otros ejemplos de sistemas que utilizan esta técnica son: P-Tango [Claypool *et al.* 99], GroupLens [Resnick *et al.* 94; Konstan *et al.* 97] y MovieLens [Good *et al.* 99; Sarwar *et al.* 01] (ver apartado 2.8 para más detalles).

2.4. Selección de contenidos

La selección de contenidos se refiere a la elección entre todos los documentos de entrada de aquellos que son más relevantes para un usuario dado, según su perfil o modelo. Para poder realizar esta selección es necesario obtener una representación de los documentos, una representación del modelo de usuario y una función que calcule la similitud entre ambas representaciones.

Para realizar la selección de contenidos existen diversos algoritmos de clasificación dependiendo de las representaciones elegidas para el modelo de usuario y los documentos. A continuación se van a describir las distintas técnicas de selección encontradas en los sistemas revisados en la bibliografía.

2.4.1. Selección basada en la fórmula del coseno del MEV

La forma más habitual de realizar la selección de contenidos cuando se utilizan términos con pesos en el MEV es mediante la aplicación de la fórmula del coseno (ecuación (2.1)) entre los vectores de pesos de términos que representan a los documentos y el vector de pesos de términos que representa al modelo de usuario. Los documentos clasificados con mayor similitud (por encima de un umbral o un número fijo de ellos) son los que son seleccionados para el usuario.

Cuando el modelo de usuario almacena varios vectores de términos se pueden combinar los resultados obtenidos para cada uno de los vectores, habitualmente seleccionando la máxima similitud vector-documento. También se puede dar mayor peso a los términos que aparecen en el título de los documentos frente a los que aparecen en el cuerpo [Nakasima&Nakamura97].

En SIFT [Yan&García-Molina95] la similitud entre un par perfil-documento es calculada utilizando la medida del coseno entre ambos vectores en el espacio vectorial. Los documentos con mayor relevancia se seleccionan para el usuario.

Otros ejemplos de sistemas que utilizan esta técnica son: Letizia [Lieberman95], PowerScout [Lieberman *et al.* 01], Amalthea [Moukas96], Pefna [Kilander *et al.* 97],

ANATAGONOMY [Sakagami&Kamba97], [Nakasima&Nakamura97], WebMate [Chen&Sycara98], PZ [Veltmann98] y [Balavanovic98b] (ver apartado 2.8 para más detalles).

2.4.2. Selección basada en el clasificador Bayes ingenuo

En este caso, para realizar la selección de los documentos que más le interesan a un usuario se utiliza la ecuación (2.6), basada en el teorema de Bayes, donde C es una de las posibles clases de evaluación (en general, relevante o no relevante), $V_{i,Ji}$ es una variable booleana con valor 1 si el documento actual contiene el atributo J_i y P(C) es la probabilidad de la clase C. La principal característica de este clasificador es que supone que los atributos son condicionalmente independientes.

$$P(C \mid V_{1,J_1},...,V_{n,J_n}) = P(C)\prod_{i=1}^{n} Q_i(C, J_i) \tag{2.6}$$

$$Q_i(C, J_i) = \frac{P(V_{i,J_i} \mid C)}{P(V_{i,J_i})} \tag{2.7}$$

Las probabilidades se pueden estimar fácilmente a partir de una colección de entrenamiento con documentos previamente clasificados.

En METIORE [Bueno01, 02] se propone una ecuación modificada (ecuación (2.8)) que da resultados similares, pero es menos restrictiva porque usa la media de los pesos de cada atributo. Esto permite obtener resultados con pocos datos de entrenamiento. Se aplica la fórmula para cada una de las clases de clasificación y la que mayor probabilidad obtenga se le asigna al documento actual.

$$P(C \mid V_{1,J_1},...,V_{n,J_n}) = P(C)\frac{\sum_{i=1}^{n} Q_i(C, J_i)}{n} \tag{2.8}$$

Además se propone una probabilidad ponderada según el tipo de atributo que se utilice (ecuación (2.9)). Donde m es el número de tipos de atributos, N_p es el número de atributos de tipo p y el factor ω_p indica la importancia que se le da al tipo de atributo p.

$$P(C \mid V_{1,J_1},...,V_{n,J_n}) = P(C)\sum_{p=1}^{m} \omega_p \frac{\sum_{i=1}^{n} Q_{pi}(C, J_{pi})}{N_p} \tag{2.9}$$

En NewsDude [Billsus&Pazzani99a, 99b] si el clasificador basado en el modelo a corto plazo no es capaz de clasificar una nueva noticia entonces se utiliza un clasificador bayesiano ingenuo donde las noticias son representadas como vectores booleanos de atributos, donde los atributos son las palabras que aparecen en los documentos. Las probabilidades del clasificador son estimadas mediante entrenamiento previo. Para clasificar una noticia como interesante se requiere que para al menos 3 palabras (atributos) se cumpla que p(palabra|interesante) > p(palabra|no interesante). Del mismo modo para clasificarla como no interesante se requiere que p(palabra|no interesante) > p(palabra|interesante).

2.4.3. Selección basada en el vecino más cercano

Un clasificador basado en el vecino más cercano (kNN => *k-nearest neighbour*)), para clasificar un nuevo documento D, selecciona los k vecinos más cercanos entre los documentos de entrenamiento, y usa las etiquetas o valores con las que estén clasificados esos documentos para predecir la etiqueta o valor del nuevo documento. Los valores de los vecinos se ponderan según la similitud entre cada vecino y el nuevo documento, donde la similitud puede ser medida por ejemplo con la distancia euclídea o con la medida del coseno, entre los vectores que representan a ambos documentos. La principal ventaja de este clasificador es que es suficiente con una noticia etiquetada para poder clasificar nuevas instancias, mientras que otros algoritmos de aprendizaje necesitan muchos más ejemplos de entrenamiento para funcionar correctamente.

En NewsDude [Billsus&Pazzani99a, 99b], cuando llega una nueva noticia el sistema intenta clasificarla utilizando los intereses a corto plazo mediante un algoritmo basado en el vecino más cercano que sigue el siguiente proceso: primero, se seleccionan las k noticias cuya cercanía respecto a la nueva noticia a clasificar, medida con la fórmula del coseno del MEV, sea mayor que un umbral mínimo t_min. La puntuación asignada a la nueva noticia por el clasificador se calcula como la media ponderada de las puntuaciones de las k noticias seleccionadas, donde la ponderación es la similitud entre la noticia seleccionada correspondiente y la noticia a clasificar. Si alguna similitud es mayor que un umbral t_max entonces se etiqueta la noticia como ya conocida y se multiplica su puntuación por un factor f << 1.0, ya que el sistema asume que el usuario ya conoce la noticia. Si no existe ninguna noticia suficientemente cercana a la noticia a clasificar (similitud > t_min), entonces ésta no se puede clasificar por el sistema a corto plazo.

El algoritmo de selección más habitual en los sistemas que utilizan filtrado colaborativo es el clasificador basado en el vecino más cercano [Goldberg *et al.* 92; Resnick *et al.* 94; Konstan *et al.* 97; Balabanovic&Shoham97; Terveen *et al.* 97; Claypool *et al.* 99; Good *et al.* 99]. Siendo los elementos, los documentos valorados por los usuarios y sus valores asociados, las valoraciones emitidas por los usuarios. De esta forma, se seleccionan aquellos documentos sobre los que los usuarios más parecidos al usuario actual, han emitido buenas valoraciones.

2.4.4. Selección basada en categorías

Como se ha comentado anteriormente, en los sistemas de personalización es muy común la utilización de temas de interés en forma de grupos de términos introducidos por los usuarios, pero la utilización de categorías es mucho más rara. De hecho, en los sistemas revisados no se ha encontrado ninguno que utilice categorías para representar modelos de usuario.

En los sistemas de categorización, la similitud entre un documento y una categoría depende de la representación elegida para la categoría, que a su vez depende de la técnica de aprendizaje seleccionada para construir el clasificador.

Una importante subclase de las técnicas basadas en aprendizaje la constituyen los clasificadores lineales o basados en perfiles, especialmente Rocchio [Hull94; Schütze *et al.* 95; Cohen&Singer96; Lewis *et al.* 96; Buenaga *et al.* 97]. Estos clasificadores almacenan de manera explícita la representación de las categorías en forma de perfiles constituidos por vectores de pesos de términos. Estos clasificadores presentan varias ventajas: son eficientes (lineales con respecto al número de términos, documentos y categorías), son fáciles de interpretar (los

vectores de pesos de términos permiten identificar fácilmente buenos predictores para las categorías), y son eficaces (muestran buenos rendimientos).

El clasificador de Rocchio es uno de los algoritmos de aprendizaje más utilizado para clasificación de texto. Se basa en una adaptación de la fórmula de Rocchio utilizada para el proceso de realimentación en recuperación de información [Rocchio71]. El clasificador de Rocchio se basa en la siguiente fórmula:

$$w_{ij} = \beta \sum_{k \in R_j} \frac{w_{ik}}{|R_j|} + \gamma \sum_{k \in R_j} \frac{w_{ik}}{|\overline{R}_j|} \qquad (2.10)$$

donde w_{ij} es el peso del término i en la categoría j, R_j es el conjunto de documentos que pertenecen a la categoría j y $\overline{R}_j$ es el conjunto de documentos que no pertenecen a la categoría j. Los parámetros β e γ (β+γ=1, β>=0, γ<=0) controlan el impacto del entrenamiento positivo y negativo, respectivamente.

En los clasificadores lineales o basados en perfil, la determinación de qué documento es relevante a cada categoría se realiza mediante la fórmula del coseno (ecuación (2.1)) entre la representación de un nuevo documento y la representación de cada una de las categorías.

2.4.5. Selección basada en estereotipos

La selección de la información se basa en una base de conocimiento de reglas que explotan la información almacenada en los estereotipos, esto es, las probabilidades asociadas a las predicciones de interés de los usuarios por los documentos.

En SeAN [Ardissono *et al.* 99a, 99b, 01] se intenta aprovechar la estructura superficial de los sistemas de noticias para personalizar la información. Las noticias están clasificadas jerárquicamente en secciones y cada una de las noticias está compuesta de distintos atributos: título, autor(es)/fuentes, un resumen, el texto del artículo, fotos y/o videos y/o audios, datos adicionales, comentarios, etc. El modelo de usuario esta basado en 4 familias de estereotipos que están fundamentados en informaciones provenientes de estadísticas anuales de la población italiana: intereses, características cognitivas, experiencia en el dominio y estilo de vida. Cada estereotipo está compuesto por una serie de atributos (edad, nivel educativo, tipo de trabajo, sexo, hobbies, receptividad, etc.), que pueden estar solapados. Inicialmente el usuario rellena un formulario con una serie de preguntas que le clasifican dentro de cada una de las familias de estereotipos. Cada estereotipo tiene dos tipos de informaciones: perfil de usuario y predicciones de interés. En el primero se almacenan una serie de probabilidades asociadas a los distintos atributos del usuario. En el segundo se almacenan las probabilidades asociadas a las predicciones de interés (alta, media, baja, nula) para cada una de las secciones.

La selección de la información se basa en un conjunto de reglas que explotan la información almacenada en los tres primeros estereotipos, esto es, las probabilidades asociadas a las predicciones de interés sobre las secciones. La selección de la publicidad se basa en la información almacenada en el último estereotipo. La selección del nivel de detalle se basa en varios atributos del modelo del usuario: experiencia del usuario (en cada sección específica), su receptividad y sus intereses. Inicialmente, lo que se le presenta al usuario son enlaces a las secciones que más le interesan al usuario, estos enlaces llevan al conjunto de noticias pertenecientes a dicha sección.

2.4.6. Selección basada en redes semánticas

Para realizar la selección se compara mediante una función de similitud la información (conceptos) extraída de los documentos a clasificar con los conceptos almacenados en la red semántica que representa el modelo del usuario. Aquellos documentos que obtengan una similitud que supere un determinado umbral de relevancia son considerados como interesantes para el usuario.

En IFTool [Asnicar *et al.* 97], para clasificar los documentos se comparan los términos que están presentes en la representación de un documento y de un modelo de usuario (redes semánticas con los intereses y los 'no intereses') y además se utiliza la información sobre las parejas de términos incluidas en el documento que ya han co-ocurrido en documentos anteriores (información representada por los arcos de la red semántica). Los documentos son clasificados como: interesantes, indiferentes o no interesantes. La utilización de la co-ocurrencia disminuye el problema de la polisemia de las palabras ya que las relaciones de co-ocurrencia permiten asociar a cada término un "contexto pragmático" que ayuda a la desambiguación de un término.

2.4.7. Selección basada en redes bayesianas

Una red bayesiana describe la distribución de probabilidad que gobierna un conjunto de atributos especificando las relaciones de probabilidad condicionada que existen entre los mismos. En el clasificador bayesiano ingenuo se supone que todos los atributos son condicionalmente independientes.

Las redes bayesianas se representan en un digrafo con n nodos A_1, ..., A_n, donde n es el número de atributos que maneja la red. Los nodos que representan atributos cuyas probabilidades son condicionalmente dependientes están unidos por arcos de tal forma que la probabilidad asociada a un nodo depende de sus antecesores. Cada nodo tiene asociado una tabla de distribución de probabilidad condicionada, que especifica su probabilidad en función de sus antecesores inmediatos en el grafo. El cálculo de la probabilidad asociada a un conjunto de valores booleanos (a_1, ..., a_n) de los atributos (A_1, ..., A_n) se puede calcular con la siguiente fórmula [Mitchell97]:

$$P(a_1,...,a_n) = \prod_{i=1}^{n} P(a_i \mid Antecesores(A_i)) \qquad (2.11)$$

En PIFT [Asnicar *et al.* 97] el clasificador está basado en dos redes bayesianas, una calcula la probabilidad de que un documento satisfaga los intereses del usuario y la otra hace un cálculo similar con los 'no intereses'. Estas redes bayesianas están constituidas por n+2 nodos S, T_1, ..., T_n, Q_I, donde n es el número de términos que aparecen en la red semántica describiendo los intereses (no intereses) de cada usuario. Se asocia cada nodo a una proposición: S se asocia con la proposición "El documento está presente en la entrada del sistema"; cada T_i está asociado con la proposición "el término t_i está presente en la representación de los intereses (no intereses) del usuario" y Q_I (Q_{NI}) está asociado con la proposición "los intereses (no intereses) del usuario se han satisfecho". Hay n arcos S->T_i y T_i->Q_I (T_i->Q_{NI}): los primeros n arcos enlazan S con cada uno de los nodos T_i y los segundos n arcos enlazan cada T_i con Q_I (Q_{NI}). Cada arco tiene asociado un valor numérico que representa la relevancia del término en la representación del documento y en la representación de los intereses (no intereses) del usuario, respectivamente. Este valor se calcula con el peso asociado a la fórmula tf

· idf [Turtle&Croft91; Callan96]. Finalmente se calcula la probabilidad de que el documento actual satisfaga los intereses ($P(Q_I|S)$) o no intereses ($P(Q_{NI}|S)$) del usuario.

En el filtrado colaborativo también se pueden utilizar redes bayesianas para realizar la selección de contenidos construyendo redes bayesianas para representar las valoraciones anteriores de los usuarios. En [Breese *et al.* 98] se utiliza una representación donde cada nodo almacena un árbol de decisión, que representa la tabla de distribución de probabilidad condicionada y cada arco almacena información del usuario. El modelo obtenido es muy pequeño, muy rápido y tan preciso como los métodos basados en el vecino más cercano.

2.4.8. Selección basada en redes neuronales

En general, la selección se realiza a través de la red neuronal, introduciendo en los nodos de entrada los atributos seleccionados para cada documento y obteniendo a la salida el valor de similitud entre el documento y el perfil de usuario representado por la red neuronal.

En [Shepherd *et al.* 02] después de una fase de entrenamiento, donde el usuario vota (interesante/no interesante) sobre un conjunto de noticias e introduce términos que definen sus intereses, el sistema puede ser utilizado para predecir nuevos intereses del usuario.

2.4.9. Selección basada en largo y corto plazo

El proceso de selección consiste en dos clasificadores que utilizan, en primer lugar, el modelo a corto plazo y en segundo lugar, el modelo a largo plazo. Los valores de similitud de los dos modelos se pueden combinar de alguna manera o se puede utilizar directamente uno de ellos. En este segundo caso, se toma el valor calculado por el corto plazo si éste puede calcular una similitud, y en caso contrario, se toma el valor calculado por el largo plazo.

En NewsDude [Billsus&Pazzani99a, 99b] se utiliza para clasificar una nueva noticia un clasificador basado en el vecino más cercano con la información almacenada en el modelo a corto plazo. Si este clasificador no puede clasificar la nueva noticia, entonces se utiliza un clasificador bayesiano ingenuo utilizando el modelo a largo plazo.

2.5. Adaptación del modelo de usuario

La adaptación del modelo de usuario es necesaria por que las necesidades de los usuarios varían con el tiempo, principalmente como efecto de la interacción con la información que reciben [Belkin97; Billsus&Pazzani00]. Por esta razón el modelo de usuario debe ser capaz de adaptarse a esos cambios de interés. Esta adaptación se realiza mediante la interacción del usuario con el sistema, a través de la cual se obtiene información para la realimentación del perfil.

Las técnicas necesarias para poder conseguir un modelado dinámico del usuario se basan en la realimentación del usuario respecto de los elementos de información que se seleccionan según su perfil. La información obtenida se utiliza para actualizar el modelo de usuario de diversas formas según sea la representación elegida. Técnicas similares se han usado con éxito para mejorar la efectividad de los sistemas de recuperación de información [Rocchio71].

Hay varias clases de técnicas de aprendizaje que se pueden utilizar para refinar los perfiles [Shepherd&Watters00]: aprendizaje directo, el usuario actualiza el perfil directamente; aprendizaje directo parcial, el usuario dice cuanto le han interesado los documentos que se le han enviado y el sistema realimenta el perfil; aprendizaje indirecto, el sistema detecta cuanto le han interesado los documentos monitorizando el comportamiento del usuario (si lo imprime, lo lee, mirando los logs, etc.) y realimenta el perfil; comunidades de filtrado, se utilizan las opiniones de una comunidad de usuarios para realizar recomendaciones.

A continuación se van a describir las distintas técnicas de adaptación encontradas en los sistemas revisados en la bibliografía.

2.5.1. Adaptación basada en el algoritmo de Rocchio o similares

Cuando se utilizan términos representados en el MEV la adaptación se realiza mediante el ajuste de los pesos de los términos o la adicción de nuevos términos extraídos de los documentos indicados como relevantes, o no relevantes, por los usuarios.

El algoritmo de Rocchio [Rocchio71] es uno de los algoritmos de aprendizaje más utilizado para clasificación de texto. Un documento se representa mediante un vector de pesos de términos, donde los pesos se calculan con tf · idf. Se puede construir un perfil inicial P = (p_1,..., p_n) como la diferencia entre la media de los ejemplos de entrenamiento positivo (μ_P) y la media de los de entrenamiento negativo (μ_N), donde α y β controlan la importancia del realimentación positiva y la negativa.

$$P = \alpha\mu_P - \beta\mu_N \tag{2.12}$$

Normalmente, el perfil obtenido se restringe a pesos positivos, por tanto, si el peso obtenido es negativo, se iguala a cero. El perfil inicial se puede utilizar para seleccionar nuevos documentos para el usuario y su realimentación (positiva o negativa) puede ser utilizada para modfiicar el perfil de usuario según la ecuación (2.13):

$$P^{new} = \gamma P^{old} + \alpha\mu_P - \beta\mu_N \tag{2.13}$$

En [Nakasima&Nakamura97] se utiliza un algoritmo similar al de Rocchio para adaptar el modelo de usuario. Se tienen en cuenta los términos de los documentos leídos y no leídos para realimentar positiva o negativamente los pesos asociados a cada uno de los términos del perfil del usuario. Además se valora adicionalmente la aparición de los términos en el título de los documentos.

El proceso seguido para realizar la realimentación consiste en calcular el valor de acceso de cada término en cada documento. Después, se suman los valores de acceso para todos los documentos, se calcula el valor de actualización de cada término y se actualiza el perfil.

Se define el valor de acceso para el término i en el documento j de la siguiente manera:

$$a\_access_{ij} = \begin{cases} \alpha \cdot (\omega \cdot b_{ij} + c_{ij}) & si\ j\ leido \\ -\beta \cdot (\omega \cdot b_{ij} + c_{ij}) & si\ j\ no\ leido \end{cases} \tag{2.14}$$

donde b_{ij} es la frecuencia del término i en el título del documento j, c_{ij} es la frecuencia del término i en el cuerpo del documento j, α es el peso de acceso, β es el peso de antiacceso y ω es el peso del título.

El valor de acceso del término i se obtiene sumando los valores de acceso en todos los documentos.

$$access_i = \sum_{j \in articulo\ filtrado} a\_access_{ij} \tag{2.15}$$

Se define el porcentaje de actualización de un término i de la siguiente manera:

$$update_i = \frac{access_i}{\max(|access_i|)} \tag{2.16}$$

El nuevo valor de interés para el término i se obtiene según la siguiente fórmula:

$$NewI_i = \begin{cases} OldI_i + (1 - OldI_i) \cdot speed \cdot update_i & (update_i \geq 0) \\ OldI_i - OldI_i \cdot speed \cdot update_i & (update_i < 0) \end{cases} \tag{2.17}$$

donde *speed* indica la velocidad de cambio del grado de interés de un término en un día. Su valor está entre 0 y 1. $NewI_i$ es el nuevo grado de interés del término i. $OldI_i$ es el antiguo grado de interés del término i.

En el artículo [Nakasima&Nakamura97] se realiza un experimento para determinar el mejor valor de los parámetros correspondientes al peso del título y el umbral, obteniéndose un valor de 3 para ambos parámetros.

En [Balabanovic98b] se realiza la adaptación del modelo de usuario a través de realimentación implícita, es decir, se ajustan los pesos de los términos del modelo de usuario ($t_i = t_i + \lambda \cdot d$) de acuerdo a los documentos sobre los que se realizan una serie de acciones en la interfaz de usuario. Por ejemplo si se mueve un documento a un tema, el valor de λ es 3, si se lee, es 0.5, si se borra es –0.3, etc.

Otros ejemplos de sistemas que utilizan técnicas similares son: Letizia [Lieberman95], Pefna [Kilander *et al.* 97], ANATAGONOMY [Sakagami&Kamba97], WebMate [Chen&Sycara98], PZ [Veltmann98], Smart Guide [Gates *et al.* 98] y [Widyantoro01] (ver apartado 2.8 para más detalles).

2.5.2. Adaptación basada en el algoritmo de Bayes ingenuo o en redes bayesianas

Las probabilidades del clasificador bayesiano ingenuo se pueden ir ajustando con los nuevos documentos juzgados por los usuarios.

En NewsDude [Billsus&Pazzani99a, 99b] se van reajustando las probabilidades del clasificador que maneja el modelo a largo plazo a través de la realimentación producida por el usuario sobre las noticias que recibe.

Las probabilidades de las redes bayesianas se pueden ir ajustando con los nuevos documentos juzgados por los usuarios de manera similar a como ocurre con el clasificador bayesiano ingenuo.

2.5.3. Adaptación basada en el vecino más cercano

La adaptación basada en el vecino más cercano se produce cuando se introducen documentos relevantes. El cálculo del vecino más cercano a la llegada de un nuevo documento será ahora distinto por la presencia de más documentos relevantes.

2.5.4. Adaptación basada en estereotipos

En general, los estereotipos suelen ser estáticos y no varían con el tiempo, ya que un usuario es asociado a un estereotipo mediante una regla de activación y las reglas de activación no cambian [Benaki *et al.* 97].

Sin embargo, en el sistema SeAN [Ardissono *et al.* 01, 99a, 99b] el modelo formado por estereotipos puede ser refinado mediante la monitorización del comportamiento del usuario a la hora de interaccionar con las noticias seleccionadas. Existe una base de conocimiento con reglas para modificar el modelo de usuario. En estas reglas los antecedentes están formados por condiciones lógicas sobre eventos y las consecuencias especifican nuevas predicciones sobre algunas características del usuario. Existen diferentes reglas según el tipo de interacción del usuario. El nuevo valor obtenido es la media entre la probabilidad generada por la predicción y la probabilidad almacenada en el modelo, haciendo que los cambios en el perfil sean suaves.

2.5.5. Adaptación basada en redes semánticas

La red semántica se puede adaptar a los cambios de interés de los usuarios mediante el ajuste de los valores asociados a los conceptos y las relaciones entre ellos, a través de la realimentación del usuario sobre los documentos recibidos.

En IFTool [Asnicar *et al.* 97] el usuario realimenta el sistema con sus juicios de interés y el sistema actualiza el modelo de usuario. Se actualizan, positiva o negativamente, los pesos asociados a los arcos de la red semántica teniendo en cuenta la frecuencia de ocurrencia de términos y la frecuencia de co-ocurrencia de parejas de términos de los documentos realimentados. Además IFTool dispone de un sistema que decrementa los pesos de los términos con el paso del tiempo para borrar términos del modelo que han podido añadirse mediante la realimentación de manera accidental. Cuando el peso de un término disminuye por debajo de un umbral, se elimina del modelo de usuario.

2.5.6. Adaptación basada en redes neuronales

La adaptación se realiza a través del algoritmo de aprendizaje que se implementa en la red.

En [Shepherd *et al.* 02] después de una fase de entrenamiento, donde el usuario vota (interesante/no interesante) sobre un conjunto de noticias e introduce términos que definen sus intereses, el sistema puede ser utilizado para predecir nuevos intereses del usuario. También se va adaptando a la realimentación del usuario a través del algoritmo de aprendizaje de la red neuronal. El sistema funciona mejor cuantas más palabras clave introduce el usuario.

2.5.7. Adaptación basada en largo y corto plazo

La adaptación de los modelos se suele realizar mediante alguna técnica de aprendizaje máquina que se suele aplicar por separado al largo y al corto plazo.

En NewsDude [Billsus&Pazzani99a, 99b] el sistema va aprendiendo (se va adaptando) con la realimentación explícita del usuario sobre las noticias que recibe (interesante, no interesante, cuéntame más). El aprendizaje se aplica tanto al modelo a corto plazo, nuevas noticias en el clasificador basado en el vecino más cercano, como en el modelo a largo plazo, cálculo de las probabilidades asociadas a palabras y clases interesante y no interesante. El aprendizaje se aplica al sistema en distintas sesiones, una sesión por día.

2.6. Presentación de resultados

La presentación de resultados consiste en, una vez seleccionados los documentos que le interesan a un usuario, generar un contenido que contenga esos documentos de manera personalizada, teniendo en cuenta las distintas preferencias indicadas por el usuario en su perfil. Los contenidos generados se plasman generalmente en un documento Web en forma de mensaje HTML que puede ser enviado por correo electrónico o visualizado como una página Web.

El objetivo de la presentación de resultados es mostrarle al usuario la información seleccionada según su perfil, pero ayudándole de alguna forma a navegar por esta información, para que el usuario pueda encontrar con la mayor facilidad posible la información que es realmente relevante para sus intereses.

2.6.1. Técnicas y métodos disponibles para la presentación de resultados

La elección más sencilla de presentación de resultados la constituye el conjunto de títulos de los documentos seleccionados para cada usuario. Por ejemplo, algunos periódicos electrónicos solamente envían por correo los títulos de las noticias seleccionadas para un usuario concreto. Algunos otros sistemas lo que muestran son los títulos y las primeras frases que aparecen en los documentos. Por ejemplo, muchos buscadores Web tienen este comportamiento. Otros buscadores muestran los títulos y las frases donde aparecen las palabras de la consulta. Estas palabras aparecen resaltadas dentro de las frases.

Además los resultados suelen aparecer ordenados por orden de similitud con los intereses del usuario. Este valor de similitud aparece asociado a cada elemento de información en forma de valor numérico o barra continua de valores.

Otras informaciones adicionales que suelen aparecer en los contenidos finales son: enlace al perfil de usuario, mecanismos para realizar realimentación, ayuda, etc.

En la mayoría de los sistemas analizados no hay una preocupación especial por la presentación de resultados, de tal forma que el principal enfoque suele consistir en mostrar simplemente los títulos de los documentos con enlaces al texto completo. En los que quizás se muestra mayor interés en la presentación es en los sistemas que utilizan realimentación implícita de los usuarios. Lógicamente para detectar los intereses de los usuarios a partir de sus acciones indirectas la interfaz del sistema debe estar mucho más cuidada que si la realimentación es explícita.

En el sistema ANATAGONOMY [Sakagami&Kamba97] se ofrece al usuario los títulos de las noticias más relevantes. Cada título tiene asociado una puntuación en forma de barra que el usuario puede modificar para mostrar su interés real sobre la noticia. Por otro lado, el usuario puede agrandar la noticia o navegar por ella mediante barras de desplazamiento, y estas acciones son interpretadas como realimentación implícita.

En [Balabanovic98b] se muestran el título y las 3 primeras frases de los documentos. El usuario puede realizar distintas operaciones sobre los documentos que recibe: mover un documento a un tema (otro panel), borrar un documento de un tema, leer un documento, asociar una estrella de oro a un documento (especialmente interesante), borrar un tema, etc. Estas acciones son interpretadas como realimentación implícita.

En METIORE [Bueno01; 02], si se activa la opción de Pedir Recomendación, se muestra una lista con los títulos de los documentos más similares al modelo de usuario junto con su valor de relevancia, o su evaluación si ya ha sido evaluado. Al seleccionar uno de ellos aparece toda la información de la que se dispone sobre el documento (título, autor, url, resumen, etc.) y opciones para que el usuario introduzca su evaluación. Si la opción activada es Búsqueda Simple los resultados se muestran de manera similar, pero la ordenación es con respecto al modelo de usuario y en caso de empate, respecto a la consulta.

En SeAN [Ardissono *et al.* 99a, 99b, 01] se generan los contenidos que se presentan a los usuarios basándose en un conjunto de probabilidades asociadas entre distintos tipos de estereotipos, las secciones a las que pertenecen las noticias y el nivel de detalle de presentación de cada una de esas noticias. Inicialmente, lo que se le presenta al usuario son enlaces a las secciones que más le interesan al usuario, estos enlaces llevan al conjunto de noticias pertenecientes a dicha sección. El nivel de detalle de estas noticias, es decir, qué partes de la noticia son mostradas (si se le presenta la noticia completa, o sólo el título, o título y autor, o título y cuerpo, con fotos o sin fotos, etc., incluso si se le manda publicidad o no) viene determinado por la información almacenada en el modelo de usuario. Además el sistema realimenta el modelo de usuario de acuerdo al comportamiento del usuario al leer las noticias (si elimina secciones o noticias, si explora noticias o secciones no seleccionadas, si elimina detalles de la noticia o si los explora, y si explora la publicidad enviada en busca de más detalles).

En [Shepherd *et al.* 02] se muestran, para cada categoría, los títulos de las noticias ordenados por relevancia, junto a cada noticia aparece las palabras clave introducidas por el editor y la posibilidad de determinar si la noticia es o no relevante para el usuario. Al seleccionar una noticia aparece su texto completo en la parte de abajo de la interfaz.

En NewsDude [Billsus&Pazzani99a, 99b] se muestra el título de las noticias más interesantes. Al seleccionar una de ellas, un sintetizador de voz lee el texto completo de la noticia. Se pueden realizar varias acciones a través de la interfaz del sistema: parar-continuar la lectura, acceder al perfil de usuario, seleccionar una categoría de donde seleccionar noticias, pasar a la siguiente noticia, realimentar el sistema o pedir explicación de la selección de la noticia.

En DailyLearner para web [Billsus&Pazzani00] se muestran los títulos de las noticias más interesantes en la parte izquierda y los textos completos en el resto de la ventana. Además los documentos aparecen ordenados por relevancia, con su valor correspondiente. Junto a los artículos completos aparece información detallada sobre su valor de relevancia, el número de usuarios que lo ha valorado, la puntuación media, la explicación de porqué ha recibido el valor de relevancia asignado y la posibilidad de realizar realimentación, tanto general sobre la noticia como particular sobre la explicación recibida.

Por otro lado, en el DailyLearner para Palm [Billsus&Pazzani00] se muestran los títulos de las 4 noticias más relevantes, ordenadas por relevancia con respecto al modelo de usuario. Adicionalmente aparece una mano con el dedo gordo hacia arriba junto a las noticias que

tienen mayor relevancia para el usuario. Existe una opción para solicitar más noticias. Cuando el usuario pulsa sobre un título aparece el primer párrafo de la noticia y la posibilidad de continuar con el siguiente párrafo. También se puede acceder a la noticia más relacionada.

En Torii [Mizarro&Tasso02] se muestran los documentos más relevantes ordenados por relevancia. La información mostrada sobre cada documento es: su título, sus autores, su relevancia en forma de barra continua, la fecha en la que fue recibido, y un enlace al documento completo.

En P-Tango [Claypool *et al.* 99], se muestra el título, sección, primeras frases y algunas veces el autor, de las noticias seleccionadas para el usuario, ordenadas por grado de interés. Al seleccionar un documento, aparece el texto completo del mismo. Además para realizar la realimentación se dispone de una barra coloreada continua que el usuario puede ajustar.

2.6.2. Generación de resúmenes

En los sistemas de personalización examinados en la bibliografía no aparece la posibilidad de mostrar al usuario un resumen, distinto de las primeras frases, de los documentos seleccionados como interesantes y mucho menos que estos resúmenes se adapten a los intereses de los usuarios descritos en sus modelos de usuario.

Por generación automática de resúmenes de texto se entiende el proceso por el cual se identifica la información sustancial proveniente de una fuente (o varias) para producir una versión abreviada destinada a un usuario particular (o grupo de usuarios) o a una determinada tarea (o tareas) [Mani01].

Existen distintos factores que afectan al proceso de generación automática de resúmenes. En [Sparck Jones99] se identificaron 3 tipos de factores: de entrada, de propósito y de salida. Entre los factores de entrada se pueden encontrar los siguientes [Alonso *et al.* 03]: estructura del documento, dominio, nivel de especialización, restricción sobre el lenguaje empleado, escala, medio, género, idioma y si se realiza sobre uno o varios documentos. Entre los factores de propósito, considerados como los más importantes en [Sparck Jones 99], se encuentran: situación (contexto de utilización), audiencia y función (sustitución del texto completo, facilitar la decisión sobre el interés del documento, ayuda para recordar el contenido completo, etc.). Por último, entre los factores de salida se encuentran [Alonso *et al.* 03]: el contenido (genérico o adaptado al usuario), el formato, el estilo (indicativo, informativo, agregativo y crítico), el proceso de construcción (extracción o abstracción), la longitud y la forma en la que se asocian al texto completo (sustitución, enlace, resaltado, etc.).

La cantidad de factores mostrados indica la dificultad en la automatización de la generación de resúmenes. Es poco probable, pues, encontrar alguna técnica o conjunto de técnicas que permita resolver el problema en todas las situaciones posibles.

Siguiendo la definición anterior [Mani01] y los diversos factores que intervienen en la generación de resúmenes, aparecen distintos sistemas de clasificación en cuanto a los tipos de resúmenes que se pueden generar. En [Hahn&Mani00] se realiza una clasificación atendiendo a su propósito, enfoque y alcance.

Atendiendo al alcance, el resumen puede limitarse a un único documento o a un conjunto de ellos que traten sobre el mismo tema.

Según su propósito o función, esto es, atendiendo al uso o tarea al que están destinados, los resúmenes se clasifican en:

- Indicativos, si el objetivo es anticipar al lector el contenido del texto y ayudarle a decidir sobre la relevancia del documento original.
- Informativos, si pretenden sustituir al texto completo incorporando toda la información nueva o trascendente.
- Agregativos, si añaden nueva información o muestran información oculta respecto al documento original.
- Críticos, si incorporan opiniones o comentarios que no aparecen en el texto original.

Finalmente, atendiendo al enfoque, podemos distinguir entre resúmenes:

- Genéricos, si recogen los temas principales del documento y van destinados a un grupo amplio de personas.
- Adaptados al usuario, si el resumen se confecciona de acuerdo a los intereses (i.e. conocimientos previos, ámbitos de interés o necesidades de información) del lector o grupo de lectores al que va dirigido.

Es razonable pensar que, sobre todo para niveles de compresión altos, un resumen que no tenga en cuenta las necesidades del usuario puede ser demasiado general como para ser útil. En concreto, en un entorno de recuperación de información, ya ha sido demostrada la superioridad de los resúmenes adaptados a la consulta realizada por el usuario [Maña *et al.* 99; Maña03]. En otros ámbitos relacionados, como el filtrado o los servicios personalizados de información, en los que el sistema puede disponer de un mayor conocimiento sobre las preferencias de los usuarios, cabría esperar una presentación más adecuada de los contenidos que simplemente los títulos o las primeras frases.

2.6.3. Técnicas y métodos disponibles para la generación de resúmenes

El proceso de generación de resúmenes tiene 3 fases [Hahn&Mani00]: analizar el texto original, determinar cuales son sus puntos más importantes y generar una salida apropiada. Las técnicas que se utilizan para generar los resúmenes se agrupan en tres tipos: técnicas de extracción de frases, técnicas basadas en relaciones discursivas y técnicas de abstracción.

2.6.3.1. Extracción de frases

Ante la variedad de tipos y dominios de los documentos disponibles, las técnicas de selección y extracción de frases resultan muy atractivas por su independencia del dominio y del idioma. El método se centra en una fase de análisis en el que se identifican los segmentos de texto (frases o párrafos, normalmente) que contienen la información más significativa. Durante esta fase se aplica un conjunto de heurísticas a cada una de las unidades de extracción. El grado de significación de cada una de ellas puede obtenerse mediante combinación lineal de los pesos resultantes de la aplicación de dichas heurísticas [Edmunson69]. Éstas pueden ser posicionales, si tienen en cuenta la posición que ocupa cada segmento dentro del documento, lingüísticas, si buscan ciertos patrones de expresiones indicativas, o estadísticas, si incluyen frecuencias de aparición de ciertas palabras.

El resumen resulta de concatenar dichos segmentos de texto en el orden en que aparecen en el documento original. La longitud del resumen puede variar según la tasa de compresión que se desee. En todo caso, de los experimentos presentados en [Morris *et al.* 92] se puede

concluir que un resumen con una tasa de compresión de entre el 20 y el 30% puede ser lo suficientemente informativo como para sustituir al documento original.

A continuación se van a describir las heurísticas más utilizadas para la extracción de frases [Hahn&Mani00; Mateo *et al.* 03]: localización, expresiones indicadoras, palabras temáticas, nombres propios, tipografía del texto y palabras especialmente importantes.

En la heurística de localización, el peso asociado depende de si la unidad de texto aparece al principio, en el medio o al final de un párrafo o del documento completo, o si aparece en determinadas partes de un documento como en la introducción o en las conclusiones. En algunos casos [Edmunson69; Kupiec *et al.* 95; Teufel&Moens97], se puntúan positivamente las frases que aparecen en los primeros y últimos párrafos. Además en [Kupiec *et al.* 95] se tiene en cuenta la posición de la frase dentro del párrafo, dando mayor peso a aquéllas que aparecen al principio y al final del mismo. También se puntúan especialmente las frases que aparecen a continuación de determinados encabezados. Estos trabajos trabajan sobre artículos científicos. Por otro lado, otros trabajos que emplean artículos periodísticos [Tombros&Sanderson98] sólo puntúan positivamente los párrafos iniciales debido a que en este tipo de textos la información importante suele aparece al principio.

Una expresión indicadora es una expresión que indica que el texto que viene a continuación es importante o significativo dentro del texto. Expresiones típicas son: "en conclusión", "en este artículo", "un resultado importante es", etc. El peso asociado también puede estar basado en términos específicos del dominio que sean especialmente significativos ("*bonus*") o negativos ("*stigma*") [Edmunson69].

El peso obtenido con la heurística de palabras temáticas se basa en la obtención de un conjunto de palabras que son especialmente representativas del contenido de un documento debido a que aparecen muy frecuentemente. Estas palabras se obtienen mediante técnicas aplicadas en la indexación de documentos: tf [Luhn58; Edmunson69; Kupiec *et al.* 95] y $tf \cdot idf$ [Teufel&Moens97]. Además se pueden tratar como palabras temáticas sólo los lemas de las palabras pertenecientes a categorías léxicas abiertas (nombres, adjetivos y verbos léxicos) [Mateo *et al.* 03]. El peso final asignado se puede calcular teniendo en cuenta grupos de palabras [Luhn58; Tombros&Sanderson98] o apariciones individuales [Edmunson69; Kupiec *et al.* 95; Teufel&Moens97]. También se puede utilizar esta heurística como filtro de frases, eliminando de las clases candidatas al resumen aquellas que obtengan un peso muy bajo [Myaeng&Jang99; Mateo *et al.* 03].

Los nombres propios suelen proporcionar información importante y su inclusión en el resumen puede aumentar la cantidad de datos facilitados al lector [Mateo *et al.* 03].

La tipografía del texto (tipo de letra, tamaño de letra, negrita, cursiva, etc.) puede utilizarse para resaltar partes especialmente importantes. Aquellas frases que contengan palabras con especial tipografía son candidatas a formar parte de un resumen [Mateo *et al.* 03].

La última heurística se basa en dar mayor peso a aquellas frases en las cuales aparecen palabras que pueden ser especialmente importantes. Por ejemplo, aquellas palabras que aparecen en el título del documento [Edmunson69; Teufel&Moens97], aquéllas que aparecen en la consulta de un usuario en un sistema de recuperación de información [Tombros&Sanderson98; Maña03] o aquéllas presentes en un modelo de usuario [Acero *et al.* 01]. También se pueden añadir nuevas palabras a partir de éstas mediante la utilización de relaciones semánticas extraídas, por ejemplo, de una base de datos léxica [Maña03].

En la mayoría de los sistemas, los pesos asociados a cada heurística, dentro de la combinación lineal que produce el valor final asociado a cada frase, son ajustados manualmente. Esta decisión está justificada por el hecho de que la contribución de cada heurística depende

del género de los textos que se estén resumiendo. Sin embargo, se puede encontrar un clasificador que determine cuales son los mejores valores de estos parámetros mediante técnicas de aprendizaje máquina [Kupiec *et al.* 95; Teufel&Moens97]. La idea es usar una colección de resúmenes generados por expertos y sus textos asociados para determinar cuales son los mejores pesos para la combinación lineal. El problema de este enfoque es la necesidad de tener un *corpus* de resúmenes y que los pesos obtenidos son dependientes del género de los documentos con los que se generen.

Los posibles problemas de inconsistencia en el resumen resultante constituyen el principal inconveniente de esta aproximación. Una forma de paliar este problema es utilizar el párrafo en lugar de la frase como unidad de extracción [Salton *et al.* 94], esperando que al proporcionar un contexto más amplio se mejoren los problemas de legibilidad.

Uno de las principales fuentes de inconsistencia son las referencias anafóricas. Una forma de resolver el problema es no incluir las frases que contengan estas referencias [Brandow *et al.* 95]. Otra forma es incluir la frase anterior [Namba&Okumura00] o las frases que resuelven la anáfora aunque no sean la inmediatamente anterior [Paice90]. El principal inconveniente de añadir estas frases es que se dejan de añadir otras frases que pueden tener mayor peso, haciendo peor el resumen.

Otra manera de resolver inconsistencias es detectar expresiones que conectan frases ("sin embargo", "por tanto", etc.) y eliminarlas si aparecen al principio de una frase y la frase anterior no forma parte del resumen [Mateo *et al.* 03].

Otro posible problema es el desequilibro, esto es, que el resumen no trate o separe los distintos subtemas o partes que pueden aparecer en el texto original. La solución más utilizada para este problema es la segmentación del texto original [Hearst97]. El resumen se puede generar utilizando las primeras frases de cada segmento [Nakao00].

Sin embargo, el método de extracción de frases tiene algunas justificaciones: aproximadamente el 80% de las frases incluidas en resúmenes manuales aparecen tal cual o con pequeñas modificaciones en el texto original [Kupiec *et al.* 95]. Por otro lado, no se encontraron diferencias significativas en un experimento en el que se pedía responder a varias preguntas, con 5 posibles respuestas, basándose en resúmenes automáticos y manuales [Morris *et al.* 92].

2.6.3.2. Relaciones discursivas

La estructura del discurso puede utilizarse para confeccionar resúmenes automáticos basándose en las relaciones semánticas que se establecen entre distintos elementos del texto original (cohesión) o utilizando las relaciones que se establecen a más alto nivel entre las frases de un texto (coherencia).

La primera técnica utiliza un grafo que representa las relaciones de cohesión, léxicas, gramaticales o semánticas, entre los elementos del texto (palabras, frases o párrafos) . Si un nodo está conectado con muchos otros significa que su presencia es especialmente importante en el texto [Mani01]. En [Skorochod'ko72] los nodos son frases y las relaciones utilizadas son las que se producen entre palabras de las frases (repetición, sinonimia, hiponimia o palabras relevantes). El resumen está formado por las frases que están más relacionadas semánticamente. Otra posibilidad es utilizar cadenas léxicas, secuencias de palabras formadas por lexemas semánticamente relacionados en un texto que tratan sobre el mismo tema [Morris&Hirst91]. Una propuesta consiste en considerar todas las cadenas léxicas posibles y elegir la que contenga mayor número de relaciones semánticas, para cada segmento del texto. Posteriormente se mezclan las cadenas dependiendo de la aparición de términos con el mis-

mo sentido. Las frases extraídas que componen el resumen se basan en la aparición de palabras que aparezcan en la cadena [Barzilay&Elhadad99].

La segunda técnica se basa en la coherencia interna del texto. La Teoría de la Estructura Retórica (RST, *Rhetorical Structure Theory*) se basa en el concepto de relación retórica entre segmentos del texto (núcleo y satélite). Se puede utilizar esta representación para determinar las unidades más relevantes e incluirlas en el resumen [Marcu00].

2.6.3.3. Abstracción

Este tipo de técnicas construyen una representación semántica del texto original y posteriormente generan el resumen a partir de dicha representación. Por tanto, el resumen puede contener frases que no aparezcan en el texto original. Esta aproximación puede conducir a crear sistemas que evitan algunos problemas, como los mencionados de inconsistencia. Sin embargo, se necesita gran cantidad de conocimiento sobre el dominio. Como consecuencia, los sistemas son poco flexibles y de difícil adaptación a otros campos de aplicación.

Este tipo de sistemas se estructuran en 3 fases: análisis, transformación y generación [Hahn&Mani00]. La primera fase construye la representación semántica. La segunda condensa la representación semántica para obtener una representación más compacta. En la tercera fase se construye el resumen en lenguaje natural a partir de la representación semántica condensada. Todas las fases necesitan una base de conocimiento para poder aplicarse, donde estas bases de conocimiento pueden ser ontologías o recursos léxicos.

Inicialmente se analiza sintácticamente el texto original y el árbol de análisis sintáctico obtenido se etiqueta con información semántica obtenida de bases de conocimiento y se convierte en estructuras conceptuales (predicados lógicos, redes semánticas, conjuntos de plantillas, etc.) que representan semánticamente al texto original. La transformación altera la representación conceptual de diversas maneras: eliminación de información redundante o irrelevante, generalización o adicción de información mediante la utilización de ontologías con conocimiento del dominio, etc. Para realizar la transformación existen varios métodos de inferencia: reglas que operan sobre asertos lógicos [Van Dijk77] u operadores que determinan patrones de actividad y conectividad en la base de conocimiento [Hahn&Reimer99]. Por último, un generador de texto traduce la representación conceptual en un resumen en lenguaje natural.

2.7. Evaluación de sistemas de personalización

Para poder medir la efectividad de los sistemas de personalización es necesario una serie de métodos que permitan medir hasta qué punto un sistema es bueno respecto a cada una de las tareas implicadas, esto es, selección, adaptación y presentación.

La mayoría de los criterios utilizados en la bibliografía para obtener los resultados de la evaluación son cuantitativos, es decir, asignan cantidades calculadas de diversas formas a las evaluaciones realizadas por el sistema, en comparación con los juicios de relevancia determinados por los usuarios. Estos criterios proceden mayoritariamente del campo de la recuperación de información. Sin embargo, cada vez más se está optando por una evaluación cualitativa basada en opiniones de los usuarios, recogidas en cuestionarios, que muestran las impresiones de los usuarios sobre la utilización del sistema en diversos aspectos. En realidad, estas dos evaluaciones son complementarias y permiten visualizar el funcionamiento del sistema

desde dos puntos de vista diferentes, desde el punto de vista del sistema y desde el punto de vista del usuario.

Para poder realizar una evaluación cuantitativa de un sistema un ingrediente fundamental son los juicios de relevancia que indican los documentos que son relevantes para cada usuario, de acuerdo a sus necesidades de información. En los sistemas de recuperación de información el conjunto de documentos relevantes para una consulta está asociado a una necesidad de información que se expresa en forma de lenguaje natural en la consulta y puede ser determinado a partir de la misma por un experto. De hecho existen colecciones de evaluación estándar (p.ej.: colecciones TREC) donde se almacenan una serie de consultas junto con los documentos relevantes asociados a cada una de ellas.

En realidad estos juicios son una aproximación de la realidad, puesto que la relevancia de los documentos dependerá del usuario que esté efectuando la consulta y del contexto concreto en el que se realice la misma. Se deberían tener en cuenta otros aspectos como la experiencia del usuario, sus preferencias, el propósito de la búsqueda, el uso de la información recibida, lo informativa que sea ésta, etc. De hecho, el concepto de relevancia ha sido ampliamente discutido en la bibliografía con distintas interpretaciones [Mizarro97].

Sin embargo, en los sistemas de personalización el conjunto de documentos relevantes es diferente para cada usuario, puesto que sus necesidades de información son diferentes, estando éstas almacenadas en su modelo de usuario. Por tanto, es necesario que cada usuario indique los juicios de relevancia sobre las noticias que recibe.

Los juicios de relevancia suelen ser binarios (relevante/no relevante), aunque en algunos casos se opta por rangos de opinión más amplios: 3 valores, 5 valores o incluso ranking de valores.

A continuación se van a presentar los métodos más habituales utilizados para la evaluación de los procesos de selección y adaptación de contenidos (apartado 2.7.1), y por otro lado, los utilizados para la presentación de resultados (apartado 2.7.2).

2.7.1. Selección y adaptación de contenidos

Para evaluar la selección se tienen en cuenta los documentos seleccionados por el sistema. El usuario juzgará estos documentos de acuerdo a sus necesidades de información, determinando si el funcionamiento del sistema es correcto.

La evaluación de la adaptación se realiza a partir del efecto que produce en la siguiente selección, por tanto, los criterios son similares a los expuestos en el párrafo anterior.

2.7.1.1. Evaluación cuantitativa

La forma de evaluar cuantitativamente una selección es comparar, de acuerdo a alguna métrica, los documentos seleccionados con los juicios de relevancia del usuario. Existen muchas métricas, procedentes de la recuperación de información, que pueden ofrecer los resultados de la evaluación en forma de curva, basada en dos valores o basadas en un único valor. Estas métricas pueden ser agrupadas en varias categorías dependiendo del tipo de relevancia y del tipo de recuperación utilizadas [VanRijsbergen79; Salton&McGill83; Mizarro01]: relevancia y recuperación binarias, relevancia binaria y recuperación en forma de ranking, relevancia y recuperación como ranking, relevancia y recuperación como ranking con valores.

2.7.1.1.1. Métricas

El primer tipo de métricas se utiliza cuando tanto la relevancia como la recuperación son binarias, es decir, los documentos son relevantes o no relevantes y el sistema recupera o no recupera cada documento. Esta dicotomía permite dividir el conjunto de todos los documentos en conjuntos de documentos recuperados y no recuperados, relevantes y no relevantes. Las métricas más conocidas son recall (R) (ecuación (2.18) y precisión (P) (ecuación (2.19)), donde el recall indica la proporción entre documentos recuperados relevantes y documentos relevantes, y la precisión indica la proporción entre los documentos recuperados relevantes y los documentos recuperados. Una medida que combina recall y precisión es F (ecuación (2.20)) [Lewis&Gale94], donde β indica la importancia relativa entre recall y precisión. Lo más habitual es que β sea igual a 1, dando igual importancia a recall y precisión. Otras medidas son: fallout, generalidad y E [VanRijsbergen79].

$$R = \frac{num.\,documentos\ recuperados\ relevantes}{num.\,documentos\ relevantes} \tag{2.18}$$

$$P = \frac{num.\,documentos\ recuperados\ relevantes}{num.\,documentos\ recuperados} \tag{2.19}$$

$$F = \frac{(\beta^2 + 1)\cdot P \cdot R}{\beta^2 \cdot P + R} \tag{2.20}$$

El segundo tipo de métricas aparece cuando la relevancia es binaria y los resultados de la recuperación se muestran en forma de ranking de documentos. El ranking viene determinado por los valores de similitud consulta-documento, determinados por el sistema que se esté evaluando. En este caso, las métricas más adecuadas son precisión para un valor fijo de recall, curva recall-precisión, recall y precisión normalizados, y longitud de búsqueda esperada.

La precisión obtenida para un valor fijo de recall supone un enfoque similar al indicado en el primer tipo de métricas: los documentos en el ranking anteriores a ese valor de recall se considerarán como documentos recuperados y los posteriores como no recuperados. Por tanto, realmente no tiene en cuenta las posiciones en el ranking de los documentos.

La curva recall-precisión se calcula en base a los valores obtenidos para la precisión en 11 niveles estándar de recall, desde 0% a 100% en intervalos de 10%. Si no se dispone de valores en esos niveles se realiza una interpolación para obtenerlos [VanRijsbergen79; Salton&McGill83]. Se puede obtener un único valor haciendo la media de la precisión para los 11 niveles de recall. Tampoco se consideran las posiciones en el ranking de los documentos entre distintos niveles de recall.

El recall y la precisión normalizados (nR y nP) [Rocchio71] miden la efectividad del ranking en función del área (en una gráfica recall o precisión versus niveles de ranking) entre la mejor solución posible (documentos relevantes en las primeras posiciones) y la solución generada por el sistema a evaluar. Estas métricas se calculan con las ecuaciones (2.21) y (2.22) donde REL es el número de documentos relevantes, $RANK_i$ representa la posición en el ranking del i-ésimo documento relevante y N es el número total de documentos en la colección. Adicionalmente, en los casos en los cuales coincide el mismo valor de relevancia para posiciones consecutivas del ranking, se toma como valor de la posición en el ranking de todos esos resultados iguales, el valor medio de las posiciones coincidentes [Salton&McGill83]. Este ajuste evita el problema de atribuir un orden aleatorio relativo a cada una de las posiciones que tienen el mismo valor de relevancia asociado, es decir, permite que la métrica sea válida cuando el orden de los documentos es un orden parcial o débil.

Estas métricas también se pueden considerar como aproximaciones al recall y precisión medios obtenidos para todos las posiciones del ranking.

$$R = 1 - \frac{\sum_{i=1}^{REL} RANK_i - \sum_{i=1}^{REL} i}{REL(N - REL)} \tag{2.21}$$

$$P = 1 - \frac{\sum_{i=1}^{REL} \log RANK_i - \sum_{i=1}^{REL} \log i}{\log N!/(N - REL)!REL!} \tag{2.22}$$

La longitud de búsqueda esperada (ESL, *expected search length*) se utiliza cuando no se puede establecer un orden total entre los documentos, si no que el orden es parcial, es decir, el ranking es de conjuntos de documentos en lugar de documentos. Por ejemplo, los documentos con mayor similitud están en el primer conjunto pero no se puede establecer un ranking entre ellos, todos tienen el mismo valor de similitud consulta-documento. La longitud de búsqueda se define como el número medio de documentos no relevantes que deben ser examinados hasta alcanzar todos los documentos relevantes [VanRijsbergen79; Salton&McGill83].

El tercer tipo de métricas aparece cuando la relevancia y la recuperación se expresan en forma de ranking de documentos. La métrica más adecuada es la medida normalizada basada en la distancia (ndpm, *normalized distance-based performance measure*) [Yao95]. Esta métrica se basa en una distancia entre dos rankings que son órdenes parciales o débiles: el ranking de recuperación y el ranking de relevancia. Se calcula sumando las distancias entre todas las posibles combinaciones de dos documentos tomados cada uno de ellos de uno de los rankings a comparar. La distancia entre dos documentos d y d', que pertenecen a distintos rankings se define como 0 si en los dos rankings d se clasifica antes de d'. Se define como 1 si en el ranking de relevancia d se clasifica antes de d' pero en el de recuperación se clasifican con la misma relevancia. Y se le asigna el valor 2 si en el ranking de relevancia d se clasifica antes de d' y en el de recuperación, d' se clasifica antes de d. La normalización se realiza sobre la máxima distancia posible. Bajo determinadas circunstancias, se pueden derivar relaciones entre esta métrica y las métricas de recall, precisión y recall normalizado [Yao95].

El cuarto tipo de métricas aparece cuando la relevancia y la recuperación se expresan en forma de ranking con valores asociados a los documentos. La métrica más adecuada es el ratio de deslizamiento [Salton&McGill83]. El valor asociado a esta métrica se calcula como la división entre la suma de los pesos de relevancia de los documentos recuperados por el sistema entre la suma de los pesos de los documentos que hubiera devuelto un sistema ideal que devolviera en primer lugar, y por orden descendente de peso, los documentos más relevantes. Se puede usar cuando los pesos asociados son 0 ó 1, resultando en una medida similar al recall normalizado o la longitud de búsqueda esperada. Una variación de esta métrica, típica de sistemas de recomendación, es el error absoluto medio, que se calcula como la diferencia media entre los pesos de relevancia encontrados por el sistema y los fijados por el usuario.

En recuperación de información el proceso de evaluación hay que repetirlo para todas las consultas. Para obtener un único valor se calcula la media de todos los valores para todas las consultas. Esta misma discusión se puede extrapolar a los procesos de selección y adaptación de contenidos teniendo en cuenta que las consultas pasan a ser los modelos de usuario y los juicios de relevancia son los propios de cada usuario.

2.7.1.1.2. Significancia estadística

Cuando se comparan dos técnicas de selección o adaptación se debe poder afirmar que una técnica es mejor que otra de manera significativa. Para poder realizar esta afirmación hay que utilizar algún test de significancia estadística [Hull93]. La mayoría de los tests estándar basados en comparaciones entre parejas de valores producirán evidencia estadística si la probabilidad de que las diferencias entre los dos conjuntos de valores no se produzcan por azar es suficientemente pequeña, habitualmente menor que 0.05. Las parejas de valores que se utilizan para comparar son los resultados obtenidos para cada técnica según la métrica utilizada.

Existen varios tests de significancia estadística que se pueden utilizar para comparar dos técnicas de selección o adaptación [Hull93]: t-test, test de signo (*sign test*) y test de Wilcoxon. En todos los casos la obtención de las medidas debe ser independiente. El t-test es un test estadístico estándar para comparar dos tablas de números, pero sólo se puede utilizar si las diferencias entre los dos conjuntos de valores están distribuidas normalmente, lo cual no es lo habitual en la comparación de técnicas de selección y adaptación. El test de signo no necesita que las distribuciones sean normales y utiliza el signo (no la magnitud) de las diferencias de los valores. El test de Wilcoxon utiliza el ranking de las diferencias (no sus magnitudes).

2.7.1.1.3. Ejemplos de evaluación cuantitativa

A continuación se muestran ejemplos de evaluación cuantitativa de alguno de los sistemas de personalización analizados.

IFTool [Asnicar *et al.* 97] se ha evaluado con 4 personas que recibieron 2000 documentos (20 cada día, durante 100 días). Los perfiles iniciales se construyen utilizando únicamente realimentación explícita. Para cada conjunto de documentos, el usuario suministra un ranking de relevancia que es utilizado para calcular el recall y la precisión normalizados, para las clasificaciones de documentos interesantes y no interesantes.

Para evaluar IfWeb [Asnicar&Tasso97] se realizó un experimento con 4 perfiles de usuario construidos inicialmente a partir de la realimentación sobre 4-6 documentos. Estos perfiles fueron utilizados para realizar una serie de 9 sesiones de funcionamiento del sistema. Después de cada sesión, se le solicitó al usuario el ranking correcto de los documentos recibidos y se comparó con el producido con el sistema. Las métricas utilizadas fueron precisión y ndpm.

En [Claypool *et al.* 99] se realiza una evaluación con 18 usuarios durante 3 semanas, 50 artículos por día, para comparar las técnicas de filtrado aisladas frente a su combinación. Se calcula el error absoluto medio entre las predicciones numéricas generadas por el sistema y las valoraciones emitidas por los usuarios.

En [Good *et al.* 99] se muestra que la combinación de técnicas de filtrado colaborativo, basadas en opiniones de los usuarios, junto a agentes de filtrado textual, produce mejores recomendaciones que cualquiera de las técnicas por separado. Estas técnicas se aplican a un subconjunto de 50 usuarios del sistema. Se construyen varios agentes de filtrado textual basado en distintas técnicas: uno basado en tf · idf modificado, otro basado en lógica inductiva, otros basados en información sobre el género de las películas, otro basado en una combinación de agentes basados en el género donde los pesos de cada agente se obtienen por regresión lineal y combinaciones de ellos. Finalmente se combinan las técnicas de filtrado textual con las de filtrado colaborativo. Una de las métricas utilizadas es el error absoluto medio.

En [Billsus&Pazzani99a, 99b] se destaca la dificultad a la hora de evaluar ese tipo de sistemas debido a varias razones. En primer lugar no es posible aplicar las metodologías estándar de aprendizaje automático (validación cruzada) debido a que el orden cronológico de los

ejemplos de entrenamiento no permite una selección aleatoria. Otro aspecto que hay que tener en cuenta es que al medir el rendimiento diario, se miden tanto los efectos de la realimentación del modelo, como el efecto del cambio de noticias. Por último, hay que resaltar que este modelo de usuario no tiene porque ser ni estático ni consistente, es decir, el mismo usuario en otro momento puede comportarse de manera distinta. El sistema es evaluado con 10 usuarios durante períodos que van desde 4 a 8 días, con 3000 noticias valoradas por los usuarios (300 de media por usuario). Las métricas utilizadas son recall, precisión y F_1.

La versión web del DailyLearner [Billsus&Pazzani00] ha sido evaluada por 150 usuarios que usaron el sistema durante más de 3 días. Las métricas utilizadas son recall, precisión y F_1. La versión PDA ha sido evaluada por 185 usuarios que usaron el sistema durante más de 10 días. Las métricas utilizadas son recall, precisión y F_1.

2.7.1.2. Evaluación cualitativa o centrada en el usuario

Para contrastar los resultados de las evaluaciones cuantitativas sobre los procesos de selección y adaptación es conveniente realizar una evaluación cualitativa, o evaluación centrada en el usuario, mediante el empleo de cuestionarios de evaluación que rellenen los distintos usuarios del sistema. Las preguntas de estos formularios sirven para ver la opinión de los usuarios sobre los distintas partes del sistema y son otra alternativa para medir la calidad del mismo.

Durante los últimos años, en la investigación alrededor de la recuperación de información, está creciendo el número de trabajos que aportan nuevas formas de estudiar este fenómeno, especialmente vinculadas a la perspectiva del usuario. La aparición de una corriente cognitiva y social, junto a una creciente aparición de trabajos de naturaleza cualitativa [Fidel93; Borrego99] suponen una buena muestra de ello, frente al enfoque tradicional basado en el sistema. Los estudios centrados en los usuarios tienen como objeto de estudio las peculiaridades contextuales e individuales de los usuarios, tanto en aspectos que influyen en las necesidades de información como en el desarrollo de la búsqueda y recuperación [Dalrymple01]. En este sentido, en [Fernández&Moya-Anegón02] se indica la existencia de dos grandes corrientes de trabajo: por un lado la sociológica y por otro la cognitiva.

Así, la corriente social, con una menor proyección, se preocupa por el estudio del contexto (institucional, cultural, etc.) y su interrelación con los usuarios de los sistemas de información. Algunos autores [Harris86; Ørom00] abogan por la validez de una serie de trabajos que apuestan por un nuevo marco teórico, intentando escapar de los paradigmas dominantes, a través de una visión unida a lo social. Esta perspectiva social se preocupa por el estudio de la formación y uso del conocimiento en un nivel socio-cultural, donde se pretende conocer el discurso en que se encuentra cada individuo

Por su parte, también aparece lo que se puede denominar la corriente cognitiva [Belkin90; Ingwersen96; Ellis96]. Esta corriente trata de potenciar lo humano frente a lo técnico típico de la perspectiva tradicional. Los principales factores estudiados son el procesamiento de la información, las modificaciones en los objetivos y estrategias de los usuarios, los aspectos afectivos y contextuales de la búsqueda de información, la influencia de las características individuales o los patrones de comportamiento. Así, es habitual encontrarse con análisis de situaciones de búsqueda en entornos "reales", empleando en diversas ocasiones métodos cualitativos con vistas a explicar cómo los usuarios gestionan su información.

Esta perspectiva parte de la consideración de la recuperación de información como un proceso interactivo en el que una parte esencial es el usuario. En este sentido, su principal pretensión es incluir en la investigación todos los factores que influyen en el uso de este tipo de sistemas. Factores que están relacionados con cuestiones como el interés que tiene el

usuario por el tema preguntado, el nivel de conocimientos que posee, con su motivación, con el modo de expresión elegido por el usuario para exponer sus necesidades o problemas de información, las técnicas y procedimientos empleados, los cambios en los estados de conocimiento, o la interacción entre estos elementos, son puntos que se convierten en nuevos objetos de estudio. Este avance no debe hacer olvidar algunos problemas, por ejemplo, se le achaca un cierto grado de ambigüedad en la delimitación de los conceptos claves como son: información, necesidad de información, significado, relevancia, satisfacción, efectividad y usuario.

En [Caro-Castro *et al.* 03] se hace un interesante repaso de gran parte de la bibliografía de la corriente cognitiva. El primer aspecto analizado es el relativo a las diferentes formas de recogida de datos, donde destaca por encima de las demás técnicas el registro de transacciones, si bien lo más habitual es combinar más de una. A continuación, aparecen los cuestionarios, las entrevistas y, finalmente, los protocolos verbales (grabación de la interacción durante la búsqueda), para acabar en los grupos de discusión y en la observación.

En segundo lugar, esta investigación deja claro que los métodos de análisis a los que más se acude son los estadísticos (donde predominan más las inferencias que la simple descripción de los resultados), aunque también son importantes los cualitativos, como por ejemplo, los análisis de contenido o la descripción densa. En cualquier caso, también es habitual utilizar ambos tipos de métodos de análisis.

Por otra parte, proponen una agrupación de las variables que aparecen en los sistemas. El primer gran grupo de variables está relacionado con el usuario. Estas variables hacen referencia a aquellos aspectos que permiten conocer la influencia de las características del usuario en las conductas de búsqueda de información y en la evaluación de los resultados. Con esto se pretende conseguir diferentes modelos de usuarios que nacen de las diferencias entre cada usuario, lo que permite diseñar y evaluar sistemas de información adaptables a los comportamientos de los diferentes grupos de comportamiento. Ejemplos de variables empleadas son las características demográficas (edad, género), las características personales, el nivel profesional, la experiencia con el sistema de información y con la búsqueda de información, el conocimiento de la materia, los modelos mentales, o los estados emocionales o afectivos.

El segundo gran grupo de variables está relacionado con el sistema. Estas variables están especialmente relacionadas con la productividad que obtiene el usuario y con el proceso de búsqueda. Las variables más importantes son las características de los sistemas, las técnicas de recuperación de información, los modelos de organización de información, la interfaz, o las características del tutorial en línea.

El tercer grupo de variables se centran en el entorno de búsqueda, que, a su vez, se subdivide en el contexto de la búsqueda de información y, por otro lado, en la definición del problema. En lo que se refiere al contexto, se suelen plantear variables relacionadas con el entorno en que el se sitúa la búsqueda de información en términos de finalidad de la búsqueda, del momento y fase de evolución de la búsqueda de información, del campo temático, y de la cantidad de información que se supone que existe. El segundo grupo de variables referidas al entorno de búsqueda es el que aborda la definición del problema de información. Aquí la clave está en la conversión de una necesidad de información en un enunciado concreto, ya sea en lo referente al nivel de definición, a la expresión o al número de búsquedas requeridas. Este grupo de variables se puede traducir en preguntas sobre el nivel de definición de la necesidad de información, sobre la forma de expresión, sobre su complejidad, o sobre el número de búsquedas por necesidad de información.

En cuarto lugar, en estos trabajos también se estudia el propio proceso de la búsqueda. Se trata de variables vinculadas a la interacción del usuario con el sistema de información, me-

diante la plasmación tanto de las acciones que se han llevado a cabo como las causas de las mismas. Así, se emplean los puntos de acceso utilizados, los términos que se han empleado en la búsqueda, las formas de combinaciones de términos de búsqueda o las órdenes y opciones del sistema utilizadas. También son relevantes la visualización de registros o documentos, los errores detectados, el tiempo empleado en cada sesión, las reacciones y emociones del usuario durante la búsqueda o la modificación del objetivo inicial de la misma.

Finalmente, y en lo que se refiere a los resultados, el análisis realizado sobre los trabajos cognitivos concluye que los modos de trabajo más empleados para medir la eficacia y la eficiencia de los sistemas son: a) el número de registros recuperados; b) la eficacia del sistema a partir de, por un lado, los juicios de relevancia por registro a partir de la relevancia temática o de la relevancia contextual (por ejemplo, la utilidad), y por otro lado, de los juicios de relevancia por búsqueda, punto donde se revisan cuestiones como la exhaustividad, la precisión, la integridad, así como la exactitud de los resultados; c) eficiencia en términos de coste de la búsqueda, de valoración subjetiva del coste de la búsqueda en función de los resultados, del valor de los resultados de búsqueda en función del tiempo empleado y del valor de búsqueda en su conjunto.

Una de las peculiaridades de la escuela cognitiva es su preocupación por generar modelos que expliquen cómo funciona la búsqueda y recuperación de información desde la perspectiva del ser humano, convirtiéndose al mismo tiempo en la base teórica para posteriores investigaciones aplicadas. Los modelos más relevantes son los siguientes [Vargas-Quesada *et al.* 02]: el modelo global de polirepresentación de Ingwersen [Ingwersen96], que hace especial hincapié en la inconsistencia del usuario para expresar su necesidad de información, el modelo episódico de Belkin [Belkin90], para el que cobran especial relevancia las interacciones que se producen entre el usuario y el sistema durante cada consulta, el modelo estratificado de Saracevic [Saracevic96], que intenta ser una mejora de las propuestas anteriores, y el modelo de retroalimentación interactiva [Spink97], donde la clave se sitúa en la influencia en la recuperación de información de los diferentes tipos de realimentación.

Los estudios asignables a esta corriente centrada en el usuario no utilizan un método unificado o un conjunto de variables similares, sino que estamos ante una línea de trabajo que está en proceso de transformación y construcción. Se trata de una perspectiva todavía joven si se compara con la tradicional en el campo de la recuperación de información. La ausencia de solidez procede en gran medida en la gran cantidad y diversidad de aspectos implicados en los procesos de búsqueda y recuperación de información así como la falta de normalización en muchos de los aspectos, procedimientos, resultados, denominación, etc.

2.7.1.2.1. Ejemplos de evaluación cualitativa

En los últimos años se ha producido un aumento en el número de estudios basados y centrados en el usuario dentro del campo de la búsqueda y recuperación de información [Beaulieu03]. Algunos de estos trabajos han dirigido su atención a sistemas de personalización de acceso a información. En [Hyldegaard&Seiden04] la evaluación que se realiza se centra en la utilidad del acceso personalizado a revistas y servicios académicos. Entre otros, su objetivo es investigar qué factores son críticos en la personalización, o qué atributos de personalización son relevantes. Se realizó la evaluación con un grupo de 14 estudiantes de doctorado, que respondieron a un cuestionario sobre tres aspectos (la amigabilidad del uso, la eficiencia en términos de tiempo, y la relevancia de la información recibida) mediante una escala del 1 (bajo) al 5 (alto). Además, el cuestionario fue seguido de una entrevista de grupo con 9 estudiantes.

En general, son frecuentes los trabajos que realizan algún tipo de evaluación cualitativa sobre sistemas de personalización de la información, en muchas ocasiones centrados en los motores de búsqueda. Por ejemplo, en [Yang&Chung04], se propone un agente inteligente que trabaja a partir de perfiles de usuarios y que aprende a partir de la realimentación, en un entorno sobre información financiera.

Hay que destacar el trabajo realizado en [Spink02], en el contexto de la interacción humana de los motores de búsqueda en web. El modelo de evaluación se fundamenta en tres elementos: uno que está vinculado al tiempo, es decir, a los movimientos y cambios que tienen lugar durante los episodios de búsqueda de información; otro que está relacionado con los episodios de búsqueda interactiva y que utiliza diferentes estudios cognitivos; y en tercer lugar, el conjunto de acciones que engloba, por ejemplo, las decisiones y juicios que se realizan durante la búsqueda de información, especialmente en torno a la consideración de la relevancia.

Como instrumentos de trabajo destacan dos cuestionarios, uno que se aplica antes del momento de la búsqueda y otro que debe ser cumplimentado con posterioridad a ésta, con los que se puede medir los cambios que se producen durante la interacción con el sistema de recuperación de información, en este caso un motor de búsqueda en Internet. También se recurre a la grabación de la interacción para recoger los comentarios que van surgiendo durante la misma. El objetivo es estudiar el comportamiento de los usuarios a partir de diferentes variables, teniendo en cuenta un modelo de trabajo previo: el problema o necesidad de información que debe ser resuelto, el proceso de búsqueda de información relacionado con el problema de información, la formulación del problema de información alrededor de una pregunta concreta, la interacción con un intermediario, la formulación de la estrategia de búsqueda con determinados términos y tácticas, actividad de búsqueda e interacciones, respuesta al usuario, evaluación del usuario sobre el impacto de la búsqueda, usabilidad. Se aplicaron los cuestionarios a 22 usuarios.

De igual forma, hay que destacar la aportación de [Pastor&Asensi99]. Los autores proponen una plantilla para evaluar el grado de satisfacción de los usuarios con respecto a los componentes gráficos del sistema, la facilidad para usar el sistema y su amigabilidad, la gestión de los contenidos, etc.

Otro trabajo en esta dirección es [Salampasis&Diamantaras02], en el que se lleva a cabo una evaluación centrada en el usuario de dos sistemas de bibliotecas digitales. Este estudio se realizó con un grupo de 24 personas que permitió extraer juicios de recall y precisión, en un contexto donde la interacción con el sistema se convierte en el eje de la investigación. De igual forma, se realiza una investigación cualitativa a partir de un cuestionario de 6 cuestiones con el que se pretende explorar si los sujetos han tenido dificultad en comprender el modelo de datos y el proceso de búsqueda, el grado de usabilidad del sistema, o qué estrategia de búsqueda era más efectiva.

2.7.2. Presentación de resultados

En principio, la evaluación de la presentación de resultados se debería realizar en función de una evaluación cualitativa donde los usuarios emitieran sus opiniones acerca de la calidad de los contenidos presentados. Lógicamente esa calidad será mayor cuanto mejor ayuden al usuario a resolver sus necesidades de información.

La mayoría de los comentarios realizados sobre la evaluación cualitativa para los procesos de selección y adaptación (apartado 2.7.1.2) es aplicable a la presentación. De hecho, en al-

gunos sistemas ya se ha citado que se realizó una evaluación sobre la interfaz o sobre la gestión de contenidos del sistema correspondiente.

En todo caso, también se puede medir cuantitativamente la presentación de resultados teniendo en cuenta la utilización de esos contenidos para resolver otra tarea de acceso a la información, o comparándolo con una presentación de resultados "ideal".

2.7.2.1. Generación de resúmenes

En particular, cuando la presentación se realiza utilizando resúmenes lo que se intenta determinar cuantitativamente es cómo de adecuado o de útil es un resumen con respecto al texto completo [Hahn&Mani00].

Desde el comienzo de la investigación en generación de resúmenes automáticos hasta el momento han ido surgiendo distintas propuestas, individuales y colectivas, para intentar consensuar los distintos aspectos relacionados con su evaluación. Por ejemplo, la conferencia SUMMAC [Mani *et al.* 98] o las conferencias DUC [Harman&Marcu01; Hahn&Harman02; Radev&Teufel03]. Sin embargo, no se ha llegado todavía a un acuerdo pleno sobre cuál es el mejor método de evaluación debido a la complejidad de la tarea.

Entre las causas que explican esta falta de acuerdo se pueden citar las siguientes [Mani01]: dificultad para llegar a un acuerdo sobre cuándo es bueno un resumen generado automáticamente (no se puede comparar únicamente con un resumen ideal generado por una persona); necesidad de utilizar personas para juzgar los resúmenes (esto hace el proceso más difícil de realizar y más difícil de repetir. Además hay problemas con el grado de coincidencia entre distintos jueces humanos); necesidad de medidas que valoren aspectos que vayan más allá de la legibilidad del resumen, por ejemplo, la relevancia de la información contenida; adaptación del resumen a las necesidades del lector y al uso al que esté destinado.

A continuación se presentarán distintos enfoques que se han presentado para el problema de la evaluación de la generación de resúmenes. Éstos pueden clasificarse, según [Sparck-Jones&Galliers96] en directos (o intrínsecos) e indirectos (o extrínsecos). Los primeros se basan en el análisis directo del resumen a través de alguna medida que permita establecer su calidad. Los segundos basan sus juicios en función de su utilidad para realizar otra tarea.

2.7.2.1.1. Evaluación directa o intrínseca

Este tipo de evaluación puede tener en cuenta [Baldwin *et al.* 00], tanto criterios de calidad, como la corrección gramatical o la cohesión del texto, como de cobertura informativa.

Los jueces humanos son imprescindibles para este tipo de evaluación, para confeccionar un resumen inicial con el que comparar o para juzgar el grado de adecuación de los resúmenes generados automáticamente. Esto supone un esfuerzo humano bastante grande, sobre todo, si tenemos en cuenta que para que los resultados sean significativos los juicios deben aplicarse a colecciones de texto grandes, como las utilizadas en las conferencias DUC. Además existe la posibilidad de que los juicios de distintos jueces no coincidan, lo cual puede invalidar los experimentos, si el desacuerdo es grande.

Existen diferentes colecciones de documentos y resúmenes que se utilizan para realizar este tipo de evaluación. Los más importantes son: conferencias DUC, TIPSTER SUMMAC, Computation and Language y RST Discourse Treebank.

Un aspecto importante a la hora de establecer los juicios de relevancia sobre los resúmenes es el grado de coincidencia entre distintos jueces. Una primera forma de medirlo es calcular el porcentaje de acuerdos, por ejemplo, mediante el índice *kappa* [Siegel&Castellan88]. En la evaluación de resúmenes se utiliza este índice para determinar el grado de concordancia

entre jueces en la elección de las frases que deberían formar parte de un resumen ideal. En [Nomoto&Matsumoto97] los jueces son estudiantes que tienen que elegir el 10% de las frases que consideran más relevantes en un *corpus* de artículos de periódico. El resultado muestra un nivel de concordancia bastante pequeño (K = 0.25).

Otra posibilidad es medir el porcentaje de coincidencia de los elementos seleccionados. En [Salton *et al.* 97] se utilizan 9 jueces para seleccionar los párrafos más importantes de una colección de 50 artículos enciclopédicos. Cada artículo es tratado por 2 jueces, con lo que se obtienen 100 resúmenes con una tasa de compresión del 20%. La media de párrafos coincidentes en los dos resúmenes es sólo del 46%.

Una tercera posibilidad lo constituye el porcentaje de acuerdo, que mide la tasa de acuerdos que se producen con la opinión mayoritaria de los jueces respecto al número total de posibles acuerdos (número de jueces por número de frases del documento). En [Jing *et al.* 98] con 5 jueces y 40 artículos de periódico se obtiene un acuerdo del 96% para una tasa de compresión del 10%. En todo caso, este valor se puede justificar por la poca longitud de los resúmenes y la utilización de textos periodísticos, donde la información más importante suele aparecer al principio del texto.

Por último, es también importante la forma en la que se explica a los jueces como deben elegir las frases, cuanto más claro esté el proceso, mayor concordancia podrá haber entre ellos.

Por otro lado, para evaluar la calidad de los resúmenes se propusieron en la conferencia DUC'02 [Hahn&Harman02] los siguientes criterios: errores en el uso de mayúsculas, orden incorrecto de las palabras, falta de concordancia en número entre sujeto y predicado, falta de componentes importantes de la frase que afecten a la claridad de la misma, fragmentos no relacionados unidos en la misma frase, omisión o uso incorrecto de artículos, pronombres con antecedentes incorrectos u omitidos, sustantivos para los que resulta imposible determinar claramente a quién o qué se refieren, posibilidad de sustituir sustantivos por pronombres, conjunciones utilizadas incorrectamente, información repetida innecesariamente y orden incorrecto de frases.

Si se obtiene una buena evaluación según los criterios anteriores podemos garantizar una buena legibilidad del resumen. Sin embargo, no se tiene en cuenta el uso que se le va a dar al resumen. Además no se garantiza que el resumen sea bueno. Es necesario utilizar otros factores que determinen la relevancia de la información que contiene o su utilidad.

El método de evaluación más utilizado para medir la relevancia del contenido de un resumen automático es la comparación con un resumen "ideal" construido manualmente. Sin embargo, la falta de concordancia entre jueces parece indicar que ese resumen no siempre existe. Por tanto, se puede dar el caso de un resumen contenga información relevante pero no se considere bueno porque no se parezca al resumen "ideal".

Las métricas utilizadas habitualmente en recuperación de información, precisión, recall y F_1 [Salton&McGill83], se han utilizado frecuentemente para evaluar la eficacia de los resúmenes automáticos con respecto al resumen ideal. Lo que se compara en este caso es el número de frases coincidentes entre el resumen automático y el ideal, dividido por el número de frases del resumen automático (Recall) o por el número de frases del resumen ideal (Precisión).

La medida más utilizada es el recall, el principal problema es que esta métrica puede dar resultados muy distintos para resúmenes semejantes en contenido, que no contengan exactamente las mismas frases.

Una primera solución consiste en dejar a criterio de los jueces qué frases del resumen automático expresan parte de información del resumen de referencia. Los jueces también determinan el porcentaje de información del resumen ideal que está cubierto en el resumen automático. Este es el criterio seguido en las conferencias DUC [Harman&Marcu01; Hahn&Harman02].

En [Donaway *et al.* 00] se proponen dos medidas alternativas a la cobertura de frases: la ordenación de frases y la similitud de contenidos. Para la primera, los jueces tienen que ordenar las frases por importancia. Posteriormente se compara con el resumen automático. El problema estriba en la dificultad para la evaluación de los jueces. La segunda medida consiste en representar ambos textos como vectores y calcular su similitud (por ejemplo, utilizando la medida del coseno [Salton&McGill83]).

Otra idea es utilizar el índice de utilidad [Radev *et al.* 00]. Se pide a los jueces una valoración, entre 1 y 10, de cada una de las frases del texto fuente. A partir de estas valoraciones se pueden construir resúmenes de cualquier longitud eligiendo las frases con mayor puntuación. El índice de utilidad se calcula dividiendo la suma de valoraciones de las frases seleccionadas por el sistema entre la suma de valoraciones de las frases del resumen "ideal". Este método tiene la ventaja de poder dar la misma relevancia a varias frases del mismo resumen, evitando el problema de la cobertura de frases, y que la evaluación de resúmenes de distintas longitudes es muy sencilla. Sin embargo, la dificultad de las decisiones de los jueces se incrementa pudiendo dar lugar a discrepancias entre ellos.

2.7.2.1.2. Evaluación indirecta o extrínseca

Los métodos de evaluación indirecta o extrínseca se basan en medir el efecto que tienen los resúmenes sobre la realización de alguna otra tarea.

En [Mani *et al.* 98b, 02] se estudió la utilidad de los resúmenes en las tareas de recuperación *ad hoc* y categorización de texto. El objetivo era determinar si la utilización de resúmenes permite un ahorro de tiempo sin pérdida de efectividad en las decisiones de relevancia.

Para la recuperación *ad hoc*, la evaluación se realizó sobre resúmenes indicativos orientados a la consulta sobre artículos de periódico. Se utilizaron 20 consultas y, para cada una de ellas se seleccionó un subconjunto de 50 documentos entre los 200 primeros recuperados por un sistema de recuperación de información estándar. La selección se realizó de manera que el número de documentos relevantes para cada consulta estuviera entre el 25 y el 75% y no hubiera documentos iguales en los subconjuntos. Por tanto, se obtuvo un *corpus* de 1000 documentos. Se utilizaron 21 jueces en la evaluación que consistía en la determinación de la relevancia de un documento dada una consulta. Los documentos que se les proporcionaron a los jueces fueron: documentos originales, resúmenes de una longitud fija (10% del número de caracteres del texto fuente), resúmenes de longitud variable y segmentos iniciales del texto fuente (10% de las frases). Este último sirve como elemento de control puesto que los documentos del experimento se caracterizan por tener mucha información relevante al comienzo de los mismos. La evaluación mostró una efectividad similar (sin diferencias significativas), en términos de F_1, cuando se utiliza el texto completo que cuando se utilizan resúmenes de longitud variable (compresión media entre 15 y 30%). Sin embargo, el ahorro de tiempo es superior al 40% cuando se utilizan los resúmenes.

Para la categorización de texto, la evaluación se centra en resúmenes genéricos, puesto que no hay consulta con la que personalizar el resumen. Se utilizaron dos grupos de 5 categorías pertenecientes a dominios excluyentes. Se seleccionaron 100 documentos para cada categoría. Se utilizaron 24 jueces en la evaluación que consistía en elegir una de entre cinco categorías predeterminadas para la cual es relevante un determinado documento o "ninguna de

ellas". Los documentos son del mismo tipo que los del experimento de recuperación *ad hoc*. Se obtuvo una efectividad similar con los dos tipos de resúmenes y con el texto completo. No obstante, sólo los resúmenes de longitud fija redujeron el tiempo, lo hicieron en un 40%.

La evaluación indirecta en tareas de acceso a la información se ha realizado en muchos trabajos. La mayoría tratan de medir las repercusiones de utilizar resúmenes en lugar de textos completos en la toma de decisiones sobre la relevancia de los documentos. La hipótesis habitual es que se produce un ahorro de tiempo sin pérdida significativa de efectividad.

En [Tombros&Sanderson98] se compara la utilidad del segmento inicial de los documentos originales con la de los resúmenes orientados a la consulta. Para ello, se utiliza un sistema de recuperación de información en el que se le muestra al usuario, junto al título de los documentos recuperados, o bien las primeras frases, o bien los resúmenes generados automáticamente. Se utilizan 50 consultas TREC y 50 documentos por consulta. Se mide la precisión, el recall y la velocidad en el proceso de decisión, las veces que el usuario accede al texto completo y la opinión subjetiva de los usuarios sobre la calidad de la ayuda (segmento inicial o resumen) suministrada. Los resultados muestran que los resúmenes orientados al usuario mejoran significativamente la eficacia de los usuarios en la tarea de recuperación ad hoc respecto a la utilización de los segmentos iniciales.

Otro enfoque diferente es el que se sigue en otros trabajos [Brandow *et al.* 95; Maña *et al.* 98, 99, 00; Nomoto&Matsumoto01]. En este caso el método de evaluación consiste en comparar la efectividad de la recuperación cuando las consultas se realizan sobre textos resumidos, en lugar de sobre la colección de documentos originales.

En [Brandow *et al.* 95] se compara la efectividad de los resúmenes automáticos con el segmento inicial de los documentos. Utiliza un *corpus* de artículos periodísticos y 12 consultas, obteniendo mayor efectividad para el segmento inicial.

En [Nomoto&Matsumoto01], basado en técnicas de agrupamiento, se utiliza un enfoque similar utilizando como sistema de control un generador de resúmenes basado en palabras clave. Se utilizan 5000 artículos de un periódico japonés y 50 consultas. Los juicios de relevancia pueden ser: total relevancia, algo relevante e irrelevante. Utilizando los dos primeros el sistema produce mayor efectividad que el de control. Si sólo se utiliza el primer nivel de relevancia, el sistema sólo es mejor para tasas de compresión de hasta el 30%. En ningún caso la efectividad alcanza la lograda con el texto completo.

En [Maña *et al.* 98, 99, 00] se utilizan 5000 documentos del *corpus* Wall Street Journal (artículos periodísticos) elegidos al azar y 50 consultas elegidas aleatoriamente con al menos un documento relevante en el conjunto seleccionado. Los experimentos llevados a cabo comparan la efectividad, mediante curvas recall-precisión, de los resultados de las búsquedas con el texto completo, con resúmenes genéricos, con resúmenes adaptados a la consulta, con resúmenes adaptados expandidos con sinónimos de WordNet y con el segmento inicial. Se han utilizado 3 tasas de compresión: 10, 15 y 30%. Los resultados muestran que los resúmenes adaptados a la consulta mejoran entre un 30 y un 40%, dependiendo de la tasa de compresión, la precisión media obtenida por las primeras frases de los textos. También mejoran a los resúmenes genéricos en una proporción similar. Finalmente, la efectividad de los resúmenes adaptados es estadísticamente comparable a la que se consigue utilizando los textos originales. Por otro lado, los experimentos también demuestran que la expansión de la consulta utilizando WordNet no mejora la efectividad de los resúmenes.

En conclusión, la evaluación indirecta sin jueces permite una evaluación mucho menos cara y mucho más extrapolable a grandes *corpora* de texto y a diversas técnicas de generación de resúmenes con diversas tasas de compresión. Las evaluaciones se pueden realizar de manera más exhaustiva y automática.

2.8. Ejemplos de sistemas de personalización

A continuación se van a mostrar un conjunto de sistemas representativos de las técnicas descritas en los apartados anteriores. Este conjunto no pretende ser exhaustivo sino mostrar una referencia de los sistemas que realizan algún tipo de personalización de contenidos.

Algunos de los sistemas se han ido describiendo parcialmente cuando se discutía sobre las técnicas de modelado de usuario, selección de contenidos, adaptación del modelo de usuario y presentación de resultados. Sin embargo, para mayor claridad, a continuación se describen todos los sistemas analizados con todas sus características.

El orden de presentación tiene que ver con el orden en el cual se describieron las distintas formas de representar modelos de usuario: términos, estereotipos, redes semánticas, redes neuronales, largo y corto plazo, filtrado colaborativo.

2.8.1. SIFT

SIFT [Yan&García-Molina95] es un ejemplo de sistema que utiliza un modelo de usuario basado en términos. En este sistema de filtrado, utilizado con grupos de noticias, cada usuario puede construir varios perfiles, donde cada perfil puede ser expresado en uno de los siguiente modelos de recuperación de información: modelo booleano o MEV.

La similitud entre un par perfil-documento es calculada utilizando la medida del coseno entre ambos vectores en el espacio vectorial. Los documentos con mayor relevancia se seleccionan para el usuario. Este es un sistema representativo de los sistemas de filtrado iniciales que no presentaban ningún tipo de adaptación.

2.8.2. Letizia y PowerScout

El sistema Letizia [Lieberman95] ayuda a la navegación sugiriendo enlaces que podrían ser de interés y que están relacionados con la página que se está visitando. Inicialmente el modelo de usuario está vacío y se va rellenando según el usuario va navegando, sin embargo, no se mantiene de unas sesiones a otras. Tanto el modelo de usuario como los documentos se representan como vectores de pesos de términos.

Se realiza una búsqueda en anchura desde la página inicial y se sugieren al usuario aquellas páginas con mayor similitud, basada en el coseno, con el perfil. Estas búsquedas y selecciones se realizan en los momentos en los que el usuario permanece inactivo en su navegación, por ejemplo, leyendo un documento. El sistema utiliza realimentación implícita para la adaptación del modelo de usuario, obteniendo el conocimiento para realizar inferencias sobre la adaptación de diversas heurísticas. Por ejemplo, si un usuario añade un documento a su lista de favoritos es que le interesa, si analiza una página y sigue algún enlace es que también le interesa, etc.

En PowerScout [Lieberman *et al.* 01], sucesor de Letizia, no se exploran documentos próximos al que el usuario está analizando sino que se generan consultas a partir de las palabras clave que se extraen del documento y se emplea un motor de búsqueda tradicional (Altavista) para obtener nuevos documentos. Además, PowerScout tiene en cuenta los diversos intereses del usuario mediante el mantenimiento de perfiles y trata de obtener, a partir de las recomendaciones que encuentra, términos que el usuario puede añadir a uno de sus perfiles.

2.8.3. Amalthea

Amalthea [Moukas96] es un sistema personalizado que trata de encontrar información interesante para los usuarios a partir de distintas fuentes. Estas fuentes pueden ser: documentos provenientes de meta-búsquedas en buscadores, documentos de fuentes periódicas o documentos de URLs seleccionadas por el usuario. La personalización se realiza a través de dos tipos de agentes: agentes de búsqueda de nueva información y agentes de filtrado. La representación del perfil y de los documentos es a través de vectores de pesos de términos, con $tf \cdot idf$ como peso, recibiendo un peso adicional los términos que aparecen en el título de los documentos. Los primeros agentes localizan los nuevos documentos y los segundos filtran la información de acuerdo a los intereses de los usuarios. El modelo de usuario inicial consiste en vectores de pesos de términos obtenidos a partir de una serie de URLs provenientes de su lista de favoritos y del histórico de navegación.

La selección se realiza mediante la fórmula del coseno entre vectores de pesos de términos. Por otro lado, el usuario puede realizar realimentación explícita sobre los elementos que recibe (escala de 5 valores). Esta realimentación afecta a la asignación de agentes de filtrado para la selección de los documentos. El usuario también puede seleccionar un conjunto de términos de un documento para realimentar su perfil de usuario. Además, el usuario puede crear nuevos agentes de filtrado seleccionando documentos, pasajes de documentos o URLs en los que está interesado.

La característica más interesante de este sistema viene dada por la posibilidad de generación de nuevos agentes de filtrado a través de cruce o mutación (algoritmos evolutivos) [Sheth&Maes93]. Los agentes poseen un valor que indica su salud dependiendo de su rendimiento (evaluable a través de la realimentación del usuario). Sólo son seleccionados para producir nuevos agentes aquellos que tengan buena salud. Además aquellos agentes que poseen mala salud son eliminados del sistema. Por otro lado, también existe un valor sobre la salud global del sistema, de tal forma que si está disminuye se acelera el mecanismo evolutivo para adaptarse a los cambios de interés del usuario. El genotipo de los agentes de filtrado es su vector de pesos de términos asociado.

2.8.4. Pefna

Pefna [Kilander *et al.* 97] es un lector de grupos de noticias que representa el modelo de usuario utilizando una lista ordenada de temas de interés definidos por el propio usuario. El usuario añade documentos representativos a cada una de los temas de interés, de tal forma que cada tema tenga uno o más documentos asociados. Se construye una representación de cada tema de interés (MEV) utilizando las palabras que aparecen en todos los documentos asociados al mismo tema.

Los nuevos documentos son comparados con cada una de los temas de interés del usuario utilizando la medida del coseno del MEV como función de similitud. Aquellos documentos que superan un umbral son seleccionados. La adaptación se realiza mediante la adicción de nuevos documentos a los temas de interés. Al añadir un nuevo documento cambia la representación del tema de interés correspondiente y por tanto, el modelo de usuario.

2.8.5. ANATAGONOMY

El sistema ANATAGONOMY [Sakagami&Kamba97], sucesor de Krakatoa [Kamba *et al.* 95; Bharat *et al.* 98], es un prototipo de un periódico Web personalizado donde los intereses de los usuarios se representan como vectores de palabras con pesos asociados. El usuario puede generar un perfil inicial suministrando una lista de palabras clave.

La selección se realiza basándose en la similitud con el vector de pesos de términos que representa el perfil del usuario [Nakamura *et al.* 95]. El sistema compara la utilidad de la realimentación implícita frente a la explícita.

En este sistema se ofrece al usuario los títulos de las noticias más relevantes. Cada título tiene asociado una puntuación en forma de barra que el usuario puede modificar para mostrar su interés real sobre la noticia. Por otro lado, el usuario puede agrandar la noticia o navegar por ella mediante barras de desplazamiento, y estas acciones son interpretadas como realimentación implícita.

La evaluación realizada con 15 usuarios durante 6 días (30 artículos por día) concluye que la realimentación explícita da mejores resultados que la implícita y por otro lado, que la combinación de ambas da resultados sólo ligeramente peores que la explícita pero disminuyendo la cantidad de trabajo (explícito) que el usuario tiene que realizar para realimentar el sistema.

2.8.6. WebMate

WebMate [Chen&Sycara98] es un agente inteligente que ayuda a los usuarios a navegar y buscar en la Web. El usuario del sistema suministra una URL que habitualmente contiene titulares de noticias. Los artículos asociados con los titulares son descargados por el sistema y comparados con el perfil del usuario para seleccionar los relevantes. El modelo de usuario almacena N vectores de pesos de términos, donde cada vector representa un interés diferente del usuario y tiene un máximo de M términos. Para construir estos vectores se utilizan documentos que el usuario ha indicado como interesantes. Se les da más peso a los términos que aparecen como título o como algún tipo de cabecera en la página HTML correspondiente al documento. En este sistema la similitud y la representación son las utilizadas en el MEV, coseno y $tf \cdot idf$.

En cuanto a la adaptación, sólo se utiliza realimentación positiva del usuario. Dado un ejemplo positivo, se construye su vector $tf \cdot idf$, si no hay N vectores se añade el vector tal cual, en caso contrario se realiza el siguiente proceso: se calcula la similitud (coseno) entre todas las parejas de vectores, incluyendo a los que ya estaban en el modelo y al nuevo vector, se combinan los vectores con mayor similitud, se ordenan los pesos del nuevo vector y se mantienen los M mayores.

2.8.7. Nakasima&Nakamura97

En [Nakashima&Nakamura97] se propone un sistema basado en agentes inteligentes aplicado a un periódico regional. El modelo de usuario almacena información "consciente" del usuario, en forma de términos con peso, e información "inconsciente", como edad, sexo, ocupación, estado civil, ciudad, etc. A partir de la información "inconsciente" se obtiene otro conjunto de términos que se almacenan en el perfil del usuario. Estos términos son actualizados por un agente inteligente dependiendo de la interacción del usuario con el sistema.

La selección se calcula utilizando una combinación de similitudes obtenidas mediante los términos "conscientes" e "inconscientes" del modelo. Estas similitudes tienen en cuenta la frecuencia de los términos del usuario en el título y en el cuerpo del documento, dando una relevancia adicional si el término aparece en el título. Los documentos que superan un cierto umbral, especificado por el sistema, son seleccionados para el usuario.

El agente encargado de la realimentación tiene en cuenta los términos de los documentos leídos y no leídos para realimentar positiva o negativamente los pesos asociados a cada uno de los términos del perfil del usuario. Además valora adicionalmente la aparición de los términos en el título de los documentos.

El proceso seguido para realizar la realimentación consiste en calcular el valor de acceso de cada término en cada documento. Después, se suman los valores de acceso para todos los documentos, se calcula el valor de actualización de cada término y se actualiza el perfil. El algoritmo se describe en el apartado 2.5.1.

2.8.8. Balabanovic98b

En [Balabanovic98b] se describe un sistema para la recomendación de noticias que el usuario puede clasificar en distintos temas mediante una interfaz de usuario que permite realizar distintas operaciones sobre los documentos que recibe: crear nuevos temas, mover un documento a un determinado tema, cambiar un documento de tema, borrar un documento, leer un documento, etc. Tanto los documentos como los usuarios se representan como vectores de pesos de términos en el MEV, utilizando tf · idf como esquema para asignar pesos a los términos. En realidad cada usuario posee un modelo de usuario compuesto por tantos vectores de términos como temas de interés tenga.

La selección se realiza mediante la fórmula del coseno del MEV entre la representación de los documentos y los perfiles de usuario. La particularidad de este sistema es que realiza la adaptación del modelo de usuario a través de realimentación implícita, es decir, ajusta los pesos de los términos del modelo de usuario ($t_i = t_i + \lambda \cdot d$) de acuerdo a los documentos sobre los que se realizan una serie de acciones en la interfaz de usuario. Por ejemplo si se mueve un documento a un tema, el valor de λ es 3, si se lee, es 0.5, si se borra es –0.3, etc.

En cuanto a la presentación de los resultados, se muestran el título y las 3 primeras frases de los documentos más relevantes.

2.8.9. PZ

El sistema PZ [Veltmann98] es un agente inteligente que genera un periódico personalizado. Está basado en una arquitectura multi-agente donde cada agente almacena un aspecto diferente de los intereses del usuario. Existen 3 tipos de agentes: agente general, que monitoriza todos los artículos y aprende a través de la realimentación del usuario, agente de secciones, que representa las secciones en las que está interesado el usuario, y agentes temáticos, que representan temas específicos en los que está interesado el usuario. Cada agente individual temático se representa como un vector de términos obtenido a través del algoritmo de Rocchio [Rocchio71]. La naturaleza dinámica de los intereses de los usuarios se tiene en cuenta mediante la asignación a cada agente temático de una cantidad inicial de "energía" que varía con el tiempo.

La selección se realiza mediante una combinación de las similitudes obtenidas por los distintos agentes que representan los intereses del usuario. Por otro lado, los vectores de términos de los agentes temáticos se adaptan mediante el algoritmo de Rocchio [Rocchio71] utilizando la realimentación explícita del usuario (interesante / no interesante). También se utiliza realimentación implícita, considerando como realimentación positiva el hecho de que un usuario seleccione y lea un artículo. Además la "energía" asociada a los agentes temáticos disminuye a lo largo del tiempo si el usuario no indica interés por el tema asociado. Si la energía disminuye por debajo de cero, se elimina el agente. Por otro lado, se pueden generar nuevos agentes temáticos dinámicamente.

2.8.10. News4U

New4U [Jones *et al.* 00] es un periódico personalizado donde las noticias pueden provenir de varias fuentes de noticias *online*. El modelo de usuario está formado por una lista de temas (subsecciones del periódico) en las que el usuario está interesado y por las fuentes de información de las que quiere recibir las noticias. El usuario puede añadir o borrar temas, asociando cada tema a un periódico concreto o a todos en general. El usuario puede introducir una serie de términos asociado a cada tema.

En este sistema se utiliza una primera selección basada en la subsección a la que pertenece la noticia, seguida por una selección basada en la similitud, a través del esquema probabilístico Okapi BM25 [Sparck Jones *et al.* 98], entre los términos asociados a las subsecciones seleccionadas por el usuario y los nuevos documentos provenientes de las fuentes de información seleccionadas por el usuario. Las noticias más relevantes para cada categoría del perfil de usuario constituyen el periódico electrónico que es ofrecido al usuario. El usuario no puede realimentar su modelo de usuario.

2.8.11. METIORE

En METIORE [Bueno01, 02] se propone un sistema de recuperación de información basado en un modelo de usuario. El enfoque de este trabajo consiste en agrupar varias actividades de búsqueda dentro de un objetivo, de tal forma que las interacciones de los usuarios sobre los resultados obtenidos para ese objetivo permiten construir un modelo de usuario que refleja los intereses del mismo. La principal aportación consiste en que la personalización permite tener en cuenta múltiples parámetros de los datos a clasificar, no sólo términos, sino también autor, fecha de publicación, etc. El esquema propuesto permite personalización sobre distintas bases de datos multimedia o documentales. También permite realizar distintos tipos de análisis globales sobre la base de datos y búsquedas clásicas con restricciones.

Los documentos se representan según una serie de parámetros que depende de la base de datos con la que se esté trabajando. Un ejemplo de conjunto de parámetros podría ser: autor, título, palabras clave, fecha de publicación, editor, editorial, url.

El modelo de usuario almacena datos personales para identificar al usuario (nombre, apellidos, correo electrónico, etc.) y una serie de objetivos que describen distintos intereses del usuario. Asociado a cada objetivo existe una descripción textual introducida por el usuario y un conjunto de valoraciones extraídas de los documentos que el usuario ha evaluado. Estas valoraciones contienen para cada parámetro utilizado, los valores que aparecen en los documentos evaluados junto al número de veces que han sido evaluados según cada tipo de evaluación posible.

Para realizar la selección de los documentos que más le interesan a un usuario se utiliza una modificación del clasificador Bayes ingenuo (ver apartado 2.4.2). Esta ecuación modificada (ecuación (2.8)) da resultados similares, pero es menos restrictiva porque usa la media de los pesos de cada atributo. Esto permite obtener resultados con pocos datos de entrenamiento. Se aplica la fórmula para cada una de las clases de clasificación y la que mayor probabilidad obtenga se le asigna al documento actual. Además se propone una probabilidad ponderada según el tipo de atributo que se utilice (ecuación (2.9)).

La adaptación del modelo de usuario se realiza mediante la información obtenida a partir de las evaluaciones del usuario. Cuando un usuario evalúa un documento con un determinado criterio, se almacena en su modelo los atributos asociados a ese documento junto con el valor asociado a la evaluación.

Hay que tener en cuenta que el usuario puede efectuar evaluaciones sobre documentos obtenidos como resultado de varios tipos de actividades de búsqueda: búsqueda simple (sólo palabras), búsqueda con restricciones, pedir recomendación (basada en modelo de usuario), ver también (documentos similares) y explotación del historial.

En cuanto a la presentación de resultados, si se activa la opción de Pedir Recomendación, se muestra una lista con los títulos de los documentos más similares al modelo de usuario junto con su valor de relevancia, o su evaluación si ya ha sido evaluado. Al seleccionar uno de ellos aparece toda la información de la que se dispone sobre el documento (título, autor, url, resumen, etc.) y opciones para que el usuario introduzca su evaluación. Si la opción activada es Búsqueda Simple los resultados se muestran de manera similar, pero la ordenación es con respecto al modelo de usuario y en caso de empate, respecto a la consulta.

2.8.12. KNOME

En KNOME[Chin89] se presenta un sistema de ayuda para el sistema operativo UNIX. La ayuda que el sistema ofrece al usuario se basa en el conocimiento que tiene el sistema del usuario. En función de las preguntas que realiza el usuario, el sistema es capaz de inferir el conocimiento de los usuarios y responder de una manera más adecuada. Se maneja un sistema de doble-estereotipo donde un conjunto de estereotipos representa el nivel de experiencia de los usuarios (novato, principiante, intermedio y experto) y otro representa el nivel de dificultad de la información (simple, usual, complejo y avanzado). KNOME almacena información, en forma de proposiciones, sobre lo que los usuarios saben y no saben. Dependiendo del diálogo con el sistema, los modelos de los usuarios van añadiendo hechos a su perfil a través de inferencias codificadas en forma de reglas.

En este sistema, para deducir si un usuario conoce un determinado hecho se utiliza un proceso de inferencia en varios pasos. Primero, se comprueba el modelo del usuario, de tal forma que si el modelo almacena que el usuario conoce o no conoce ese hecho entonces la respuesta es, verdadero o falso, respectivamente. Si no es así, entonces se comprueba el estereotipo, si los usuarios pertenecientes al mismo conocen o no conocen el hecho, de nuevo la respuesta es verdadero o falso, respectivamente. Si tampoco es factible, la decisión se basa en el nivel de dificultad del hecho. Si el estereotipo conoce todos los hechos de ese nivel de dificultad, entonces la respuesta es verdadero, si conoce la mayoría, la respuesta es probablemente, si conoce unos pocos, la respuesta es probablemente no. Si este último proceso falla, la respuesta es incierto.

2.8.13. SeAN

En SeAN [Ardissono *et al.* 99a, 99b, 01] se intenta aprovechar la estructura superficial de los sistemas de noticias para personalizar la información. Las noticias están clasificadas jerárquicamente en secciones y cada una de las noticias está compuesta de distintos atributos: título, autor(es)/fuentes, un resumen, el texto del artículo, fotos y/o videos y/o audios, datos adicionales, comentarios, etc. El modelo de usuario esta basado en 4 familias de estereotipos que están fundamentados en informaciones provenientes de estadísticas anuales de la población italiana: intereses, características cognitivas, experiencia en el dominio y estilo de vida. Cada estereotipo está compuesto por una serie de atributos (edad, nivel educativo, tipo de trabajo, sexo, hobbies, receptividad, etc.), que pueden estar solapados. Inicialmente el usuario rellena un formulario con una serie de preguntas que le clasifican dentro de cada una de las familias de estereotipos. Cada estereotipo tiene dos tipos de informaciones: perfil de usuario y predicciones de interés. En el primero se almacenan una serie de probabilidades asociadas a los distintos atributos del usuario. En el segundo se almacenan las probabilidades asociadas a las predicciones de interés (alta, media, baja, nula) para cada una de las secciones.

La selección de la información se basa en un conjunto de reglas que explotan la información almacenada en los tres primeros estereotipos, esto es, las probabilidades asociadas a las predicciones de interés sobre las secciones. La selección de la publicidad se basa en la información almacenada en el último estereotipo. La selección del nivel de detalle se basa en varios atributos del modelo del usuario: experiencia del usuario (en cada sección específica), su receptividad y sus intereses.

El modelo inicial del usuario es genérico e impreciso, debido al limitado conjunto de preguntas que se le hacen al usuario. Sin embargo, el modelo puede ser refinado mediante la monitorización del comportamiento del usuario a la hora de interaccionar con las noticias seleccionadas. Existe una base de conocimiento con reglas para modificar el modelo de usuario. En estas reglas los antecedentes están formados por condiciones lógicas sobre eventos y las consecuencias especifican nuevas predicciones sobre algunas características del usuario. Existen diferentes reglas según el tipo de interacción del usuario. El nuevo valor obtenido es la media entre la probabilidad generada por la predicción y la probabilidad almacenada en el modelo, haciendo que los cambios en el perfil sean suaves.

Se generan los contenidos que se presentan a los usuarios basándose en un conjunto de probabilidades asociadas entre distintos tipos de estereotipos, las secciones a las que pertenecen las noticias y el nivel de detalle de presentación de cada una de esas noticias. Inicialmente, lo que se le presenta al usuario son enlaces a las secciones que más le interesan al usuario, estos enlaces llevan al conjunto de noticias pertenecientes a dicha sección. El nivel de detalle de estas noticias, es decir, qué partes de la noticia son mostradas (si se le presenta la noticia completa, o sólo el título, o título y autor, o título y cuerpo, con fotos o sin fotos, etc., incluso si se le manda publicidad o no) viene determinado por la información almacenada en el modelo de usuario. Además el sistema realimenta el modelo de usuario de acuerdo al comportamiento del usuario al leer las noticias (si elimina secciones o noticias, si explora noticias o secciones no seleccionadas, si elimina detalles de la noticia o si los explora, y si explora la publicidad enviada en busca de más detalles).

2.8.14. Browse

Browse [Jennings&Higuchi92] es un lector de grupos de noticias que utiliza redes neuronales para representar los intereses del usuario. La red neuronal almacena asociaciones entre pare-

jas de palabras (no necesariamente adyacentes en el texto) de tal manera que llega a ser sensible a probabilidades de co-ocurrencia. Cada nodo está asociado a una palabra que aparece en artículos que ha leído el usuario y tiene una energía que determina lo importante que es esa palabra para el usuario. Por otro lado, los enlaces representan el grado de asociación entre palabras, en el sentido de co-ocurrencia y también tienen un peso que indica la importancia para el usuario. Las primeras 300 palabras de los artículos (filtradas de palabras vacías) representan los atributos del documento y son las que sirven para construir/ajustar la red.

El sistema clasifica los documentos en uno de los siguientes 3 modos: todos los documentos, artículos seleccionados ordenados por relevancia y búsqueda de términos aumentada. El primer modo simplemente muestra todos los documentos en orden cronológico. El segundo utiliza la red neuronal para clasificar los documentos y filtrarlos de acuerdo a un umbral fijado por el usuario. En el tercer modo el usuario puede introducir una serie de términos que incrementarán o disminuirán la capacidad de aceptación de un documento.

La clasificación de la red neuronal se basa en los atributos extraídos de cada documento y en la activación de los nodos correspondientes a esos atributos y en los que están conectados con estos. La suma de las energías y conexiones de los nodos activos determina el valor de similitud entre artículo y modelo de usuario. Cada mensaje que se presenta al usuario se clasifica como aceptado (leído) o no aceptado (no leído). Finalmente, los mensajes aceptados o rechazados actualizan la red neuronal que representa los intereses del usuario.

2.8.15. Shepherd *et al.* 02

En [Shepherd *et al.* 02] se propone un sistema adaptativo de filtrado de noticias. El sistema está basado en un modelo de usuario que integra estereotipos y redes neuronales. Los estereotipos están basados en las secciones y sub-secciones de un periódico, además de incluir palabras claves introducidas por el editor y palabras clave introducidas por el usuario. La red neuronal encapsula información sobre los estereotipos para implementar perfiles de usuario adaptativos que reflejan los cambios de interés de un usuario. Posee 114 nodos de entrada: 79 para cada una de las sub-secciones, 2 nodos por cada una de las secciones indicando la presencia o ausencia de palabras claves del editor y del usuario y un nodo que representa si el usuario ha encontrado interesante el artículo. La capa oculta dispone de 17 nodos correspondientes a las secciones que se enlazan con los nodos de entrada que corresponden a subsecciones de cada sección. Por otro lado, hay un único nodo de salida que representa dos posibles estados con respecto al artículo: interesante o no interesante.

Después de una fase de entrenamiento, donde el usuario vota (interesante/no interesante) sobre un conjunto de noticias e introduce términos que definen sus intereses, el sistema puede ser utilizado para predecir nuevos intereses del usuario.

En cuanto a la presentación de resultados, se muestran, para cada categoría, los títulos de las noticias ordenados por relevancia, junto a cada noticia aparece las palabras clave introducidas por el editor y la posibilidad de determinar si la noticia es o no relevante para el usuario. Al seleccionar una noticia aparece su texto completo en la parte de abajo de la interfaz.

2.8.16. IFTool

En IFTool [Asnicar *et al.* 97] se describe un sistema de filtrado basado en un modelo de usuario que describe los "intereses" y "no intereses" de un usuario. Más concretamente, está formado por dos redes semánticas, una que representa los "intereses" del usuario y otra que

representa los "no intereses", es decir, información sobre la que no está interesado el usuario. Cada red semántica contiene nodos que se corresponden con términos (conceptos) encontrados en los documentos y los arcos unen términos que co-ocurren en el mismo documento. Cada nodo tiene un peso que puede ser positivo o negativo dependiendo de si se ha extraído de un documento juzgado como interesante o como no interesante. Cada arco está caracterizado con un peso que indica la frecuencia de la co-ocurrencia de los términos en los documentos previamente analizados. La utilización de la co-ocurrencia disminuye el problema de la polisemia de las palabras ya que las relaciones de co-ocurrencia permiten asociar a cada término un "contexto pragmático" que ayuda a la desambiguación de un término.

Para clasificar los documentos se comparan los términos que están presentes en la representación de un documento y de un modelo de usuario (redes semánticas con los intereses y los 'no intereses') y además se utiliza la información sobre las parejas de términos incluidas en el documento que ya han co-ocurrido en documentos anteriores (información representada por los arcos de la red semántica). Los documentos son clasificados como: interesantes, indiferentes o no interesantes.

En cuanto a la adaptación, el usuario realimenta el sistema con sus juicios de interés y el sistema actualiza el modelo de usuario. Se actualizan, positiva o negativamente, los pesos asociados a los arcos de la red semántica teniendo en cuenta la frecuencia de ocurrencia de términos y la frecuencia de co-ocurrencia de parejas de términos de los documentos realimentados. Además IFTool dispone de un sistema que decrementa los pesos de los términos con el paso del tiempo para borrar términos del modelo que han podido añadirse mediante la realimentación de manera accidental. Cuando el peso de un término disminuye por debajo de un umbral, se elimina del modelo de usuario.

Este sistema se ha evaluado con 4 personas recibiendo 2000 documentos (20 cada día, durante 100 días). Los perfiles iniciales se construyen utilizando únicamente realimentación explícita. Para cada conjunto de documentos, el usuario suministra un ranking de relevancia que es utilizado para calcular el recall y la precisión normalizados, para las clasificaciones de documentos interesantes y no interesantes. Los resultados muestran buenas capacidades de aprendizaje ya que se alcanza un 80% de precisión normalizada después de 8 sesiones llegándose a un valor final del 92% tras 100 sesiones.

2.8.17. ifWeb

En ifWeb[Asnicar&Tasso97] se propone un sistema que colabora con un navegador para refinar el funcionamiento del mismo aplicando un modelo del usuario. El sistema puede operar de dos modos distintos: como ayuda a la navegación, partiendo del documento que se esté explorando, recupera y clasifica otros documentos y presenta al usuario la estructura de las relaciones entre esos documentos y el que se está explorando; o como buscador, recupera y clasifica documentos de forma autónoma y muestra al usuario aquellos que puedan resultarle más interesantes. Los mecanismos de representación del usuario, selección y adaptación están basado en IFTool.

En este sistema los resultados se muestran como títulos de páginas enlazados en una estructura de árbol, donde los arcos corresponden a hiperenlaces. Existen varios iconos para representar la relevancia de los resultados: +, para interesante, -, para no interesante, =, para indiferente, ?, para no realizado, STOP, para indicar que se ha decidido no continuar el análisis a partir de ese documento. Si el usuario selecciona un documento este aparece en otra

ventana del navegador, donde se dispone de dos botones para realizar realimentación explícita (interesante / no interesante).

2.8.18. PIFT

En PIFT [Asnicar *et al.* 97] se utiliza un modelo de usuario análogo al de IFTool, esto es, se utilizan dos redes semánticas, una para los "intereses" del usuario, y otra para los "no intereses". Cada nodo se asocia con un término y los nodos unen términos que co-ocurren en el mismo documento, pero dentro de una distancia limitada de tamaño m, llamada ventana contextual. Estas ventanas tratan de eliminar co-ocurrencias entre términos muy alejados que no tienen porque representar una información significativa [Xu&Croft96]. Se asocia un peso a cada nodo utilizando la medida tf · idf, y un peso a cada arco que depende de la frecuencia de co-ocurrencia de los términos que enlaza.

El clasificador está basado en dos redes bayesianas, una calcula la probabilidad de que un documento satisfaga los intereses del usuario y la otra hace un cálculo similar con los 'no intereses' (ver apartado 2.4.7). El valor obtenido mediante la combinación de ambas redes se utiliza para calcular las relevancias asociadas a los documentos.

Los resultados obtenidos con PIFT son ligeramente peores que los obtenidos con IFTool.

2.8.19. NewsDude y DailyLearner

En NewsDude [Billsus&Pazzani99a, 99b] se construye un programa personalizado de noticias basado en un sintetizador de voz (en teoría para una radio de un coche). Además de representar por separado los intereses a corto y largo plazo, ambos de manera dinámica, tiene en cuenta las noticias que ya han sido presentadas al usuario para evitar presentar la misma información dos veces.

Se utilizan las 100 últimas noticias sobre las que el usuario ha emitido algún tipo de juicio para representar los intereses a corto plazo del usuario. Estas noticias se representan como vectores de pesos de términos y tienen asociada una puntuación, entre 0 y 1, que se obtiene de la valoración de los usuarios de la siguiente forma: si el usuario elige "no interesante" => 0.3 * pl, si elige "interesante" => 0.7 + 0.3 * pl, si pregunta por más información => 1.0 (siendo pl la proporción del artículo que el usuario ha oído).

Por otro lado los intereses a largo plazo sirven para modelar las preferencias generales del usuario y se utilizan cuando un nuevo artículo no puede ser clasificado por el modelo a corto plazo. En este caso, se representan las noticias como vectores de booleanos, donde cada posición del vector está asociada a un determinada palabra (característica). Estos términos se eligen manualmente como buenos indicadores de las noticias que aparecen más comúnmente. Se seleccionaron aproximadamente 200 palabras pertenecientes a distintos temas, como atributos para un clasificador bayesiano ingenuo. También se almacena la opinión del usuario ("interesante" o "no interesante") sobre las noticias.

En este sistema cuando llega una nueva noticia el sistema intenta clasificarla utilizando los intereses a corto plazo mediante un algoritmo basado en el vecino más cercano (ver apartado 2.4.3) que sigue el siguiente proceso: primero, se seleccionan las k noticias cuya cercanía respecto a la nueva noticia a clasificar, medida con la fórmula del coseno del MEV, sea mayor que un umbral mínimo t_min. La puntuación asignada a la nueva noticia por el clasificador

se calcula como la media ponderada de las puntuaciones de las k noticias seleccionadas, donde la ponderación es la similitud entre la noticia seleccionada correspondiente y la noticia a clasificar. Si alguna similitud es mayor que un umbral t_max entonces se etiqueta la noticia como ya conocida y se multiplica su puntuación por un factor f << 1.0, ya que el sistema asume que el usuario ya conoce la noticia. Si no existe ninguna noticia suficientemente cercana a la noticia a clasificar (similitud > t_min), entonces ésta no se puede clasificar por el sistema a corto plazo.

Si el primer sistema no es capaz de clasificar la nueva noticia entonces se utiliza un clasificador bayesiano ingenuo donde las noticias son representadas como vectores booleanos de atributos. La puntuación final asociada a una nueva noticia perteneciente a una clase j dado un vector de palabras (p_1, ..., p_n), consideradas como atributos del clasificador, es proporcional a:

$$p(clase_j)\prod_{i=1}^{n} p(p_i \mid clase_j) \tag{2.23}$$

Las probabilidades del clasificador son estimadas mediante entrenamiento previo. Para clasificar una noticia como interesante se requiere que para al menos 3 palabras (atributos) se cumpla que p(palabra|interesante) > p(palabra|no interesante). Del mismo modo para clasificarla como no interesante se requiere que p(palabra|no interesante) > p(palabra|interesante). Si el sistema tampoco es capaz de clasificarla, se le asigna una puntuación por defecto (0.3) que debe ser mayor que la relevancia de cualquier noticia no interesante.

Además el sistema puede generar una explicación, si el usuario lo solicita, de porqué se ha seleccionado una noticia, para ello se basa en si se ha clasificado con el primer o el segundo sistema, no se ha podido clasificar o el sistema determina que ya se ha leído.

El sistema va aprendiendo (se va adaptando) con la realimentación explícita del usuario sobre las noticias que recibe (interesante, no interesante, cuéntame más). El aprendizaje se aplica tanto al modelo a corto plazo, nuevas noticias en el clasificador basado en el vecino más cercano, como en el modelo a largo plazo, cálculo de las probabilidades asociadas a palabras y clases interesante y no interesante. El aprendizaje se aplica al sistema en distintas sesiones, una sesión por día.

En cuanto a la presentación de resultados, se muestra el título de las noticias más interesantes. Al seleccionar una de ellas, un sintetizador de voz lee el texto completo de la noticia. Se pueden realizar varias acciones a través de la interfaz del sistema: parar-continuar la lectura, acceder al perfil de usuario, seleccionar una categoría de donde seleccionar noticias, pasar a la siguiente noticia, realimentar el sistema o pedir explicación de la selección de la noticia.

En este trabajo se destaca la dificultad a la hora de evaluar ese tipo de sistemas debido a varias razones. En primer lugar no es posible aplicar las metodologías estándar de aprendizaje automático (validación cruzada) debido a que el orden cronológico de los ejemplos de entrenamiento no permite una selección aleatoria. Otro aspecto que hay que tener en cuenta es que al medir el rendimiento diario, se miden tanto los efectos de la realimentación del modelo, como el efecto del cambio de noticias. Por último, hay que resaltar que este modelo de usuario no tiene porque ser ni estático ni consistente, es decir, el mismo usuario en otro momento puede comportarse de manera distinta.

El sistema es evaluado con 10 usuarios durante períodos que van desde 4 a 8 días, con 3000 noticias valoradas por los usuarios (300 de media por usuario). Las métricas utilizadas son recall, precisión y F_1. Los resultados muestran que el rendimiento crece rápidamente durante las 3 primeras sesiones. En los usuarios que utilizaron el sistema durante más de 3

días se observa que después del crecimiento inicial el sistema fluctúa debido al cambio de noticias diario. Los resultados también muestran que el modelo híbrido es el que mejores resultados da, seguido por el largo plazo. Por otro lado, si se elimina la posibilidad de que el sistema detecte una noticia como ya leída entonces los resultados obtenidos son peores.

En [Billsus&Pazzani00] se extienden las ideas presentadas en NewsDude a un marco cliente-servidor para el acceso a noticias de manera adaptativa. El centro del marco es el servidor de información adaptativo, el cual utiliza un algoritmo de aprendizaje basado en multi-estrategia, diseñado para manejar modelos de usuario a corto y largo plazo. Se presentan dos diferentes versiones de un agente para el acceso a noticias de manera adaptativa, Daily Learner. El primero a través de una interfaz Web y el segundo orientado a dispositivos de información inalámbricos como PDAs o móviles.

La principal diferencia respecto a NewsDude es la elección de los atributos para el clasificador bayesiano ingenuo, que en este caso se realiza automáticamente generándose un conjunto de atributos diferente para cada categoría de noticias. Primero se seleccionan las n (10) palabras más informativas (mayor tf · idf) de cada documento, después se generan los conjuntos de atributos utilizando las palabras más informativas que más aparecen en un conjunto de m (10000) documentos de la categoría correspondiente, finalmente se seleccionan las k (150) palabras con mayor frecuencia como atributos de la categoría.

El mecanismo de selección es el mismo que en NewsDude. Por otro lado, el mecanismo de realimentación explícita del cliente Web es similar al de NewsDude, mientras que en el cliente para PDA se utiliza realimentación implícita: se fija una puntuación inicial (0.8) si el usuario pincha en el título de una noticia y se va incrementando dicha relevancia según se va solicitando más información. Si el usuario descarga la noticia completa la puntuación es 1. Si el usuario no selecciona una noticia se interpreta como realimentación negativa (no interesante) pero no se le asigna el valor por defecto sino que se le asigna el valor de la predicción menos una constante.

En DailyLearner para web se muestran los títulos de las noticias más interesantes en la parte izquierda y los textos completos en el resto de la ventana. Además los documentos aparecen ordenados por relevancia, con su valor correspondiente. Junto a los artículos completos aparece información detallada sobre su valor de relevancia, el número de usuarios que lo ha valorado, la puntuación media, la explicación de porqué ha recibido el valor de relevancia asignado y la posibilidad de realizar realimentación, tanto general sobre la noticia como particular sobre la explicación recibida.

Por otro lado, en el DailyLearner para Palm se muestran los títulos de las 4 noticias más relevantes, ordenadas por relevancia con respecto al modelo de usuario. Adicionalmente aparece una mano con el dedo gordo hacia arriba junto a las noticias que tienen mayor relevancia para el usuario. Existe una opción para solicitar más noticias. Cuando el usuario pulsa sobre un título aparece el primer párrafo de la noticia y la posibilidad de continuar con el siguiente párrafo. También se puede acceder a la noticia más relacionada.

La versión web del DailyLearner ha sido evaluada por 150 usuarios que usaron el sistema durante más de 3 días. Las métricas utilizadas son recall, precisión y F_1. De manera similar a los experimentos con NewDude, los resultados muestran que el rendimiento crece rápidamente durante las 3 primeras sesiones. En los usuarios que utilizaron el sistema durante más de 3 días se observa que después del crecimiento inicial el sistema fluctúa debido al cambio de noticias diario. Los resultados también muestran que el modelo híbrido es el que mejores resultados da, seguido por el corto plazo.

La versión PDA ha sido evaluada por 185 usuarios que usaron el sistema durante más de 10 días. Las métricas utilizadas son recall, precisión y F_1. De manera similar, los resultados

muestran que el rendimiento crece rápidamente durante las 5 primeras sesiones y luego fluctúa debido al cambio de noticias diario. Los resultados también muestran que el modelo híbrido es el que mejores resultados da, seguido por el corto plazo.

Los valores de precisión y recall son mucho más bajos en la versión PDA (32.42 y 29.26) que en la versión Web (71.72 y 55.55). Esto es debido a que la versión web utiliza realimentación explícita mientras que la versión PDA utiliza realimentación implícita. Como resultado adicional se pudo observar que el número de noticias etiquetadas como interesantes varía mucho entre las dos versiones (aproximadamente 58% en la versión web, por 18.3% para la versión PDA).

2.8.20. Widyantoro01

En [Widyantoro01] se utiliza una combinación de largo y corto plazo basada en 3 descriptores: uno para el largo plazo y dos para el corto plazo (uno para intereses positivos y otro para intereses "negativos"). Cada descriptor es un conjunto de pares atributo-valor donde el atributo representa un término y su valor (peso tf · idf) indica la importancia del mismo, es decir, vectores de pesos de términos al estilo del MEV. Además se almacena un peso asociado a los intereses positivos y otro asociado a los negativos. La obtención de estos descriptores se realiza a partir de la realimentación del usuario sobre el sistema.

Este esquema se puede utilizar para representar un único interés de un usuario en un tema o categoría, o se puede extender para representar múltiples intereses temáticos, cada uno de ellos representado por los 3 descriptores descritos anteriormente.

La selección se realiza calculando la similitud, mediante la fórmula del coseno (ecuación (2.1)), entre la representación de los documentos (MEV) y los vectores de pesos de términos que representan los 3 descriptores que constituyen el modelo del usuario. La relevancia obtenida para el corto plazo se obtiene como la resta ponderada (peso asociado) de la similitud obtenida para los intereses positivos menos la obtenida para los negativos. La relevancia final se obtiene como la suma de las similitudes obtenidas para el corto y largo plazo.

Si lo que tenemos son múltiples intereses en el modelo de usuario, entonces la similitud entre el nuevo documento y el modelo de usuario se calcula como la máxima similitud entre el nuevo documento y cada uno de los intereses del modelo.

La adaptación se realiza documento a documento, no se realiza al final de una sesión, a través de la realimentación explícita del usuario. La adaptación del modelo a largo plazo se realiza mediante el algoritmo de Rocchio [Rocchio71] utilizando los documentos sobre los que el usuario haya realimentado tanto positiva como negativamente. Se justifica la utilización de este algoritmo porque su regla de actualización de pesos produce cambios graduales de interés. La adaptación del modelo a corto plazo afecta a los dos descriptores que representan los intereses y los no intereses. Al recibir un documento con realimentación positiva se ajusta el descriptor de los intereses del usuario sumando los pesos del descriptor con los pesos del documento de manera ponderada, esto es, se multiplican por un factor α los pesos del nuevo documento y por un factor 1-α, los del descriptor. De manera similar se ajusta el peso asociado a los no intereses. Y por otro lado, se ajusta el peso asociado a los no intereses restando la similitud, ponderada con α, de los no intereses con respecto al nuevo documento. Este valor de α permite ajustar el cambio producido por la realimentación. Si la realimentación es negativa los ajustes en los no intereses se realizan de manera similar.

Si el modelo de usuario almacena múltiples intereses temáticos la adaptación se produce según el proceso descrito a continuación. En primer lugar, se calcula la máxima similitud entre cada interés temático y el nuevo documento, siendo la similitud entre el nuevo documento y un interés temático calculada como la máxima similitud entre documento y largo plazo, corto plazo positivo y corto plazo negativo. En segundo lugar, se crea un nuevo interés temático si esta similitud es menor que un umbral θ y el número de intereses temáticos es menor que una constante M, si no es menor se pueden seguir dos tácticas para actualizar los intereses temáticos: actualizar el interés que haya obtenido la mayor similitud o actualizar todos aquellos intereses cuya similitud supere el umbral θ.

Se realiza una evaluación con un subconjunto de la colección Reuters-21578 1.0 seleccionando los 12902 artículos que tienen asignadas una o más categorías. Los resultados muestran que los modelos basados en una representación con 3 descriptores ofrecen mejores resultados que los modelos con un único descriptor, tanto en términos de velocidad de recuperación al producirse cambios de interés sencillos, como en términos de precisión asintótica cuando se tienen que manejar intereses temáticos múltiples.

2.8.21. Torii

En Torii [Mizarro&Tasso02] se distinguen dos tipos de personalización, persistente (o a largo plazo) y efímera (o a corto plazo), en un portal de publicaciones académicas. En el primer tipo, el perfil de usuario se desarrolla incrementalmente a lo largo del tiempo y se va almacenando para ser usado en sesiones sucesivas. En el segundo tipo, la información utilizada para obtener el perfil de usuario se obtiene únicamente de la sesión actual y se utiliza inmediatamente en la misma sesión para realizar algún tipo de personalización. Al final de la sesión actual este perfil no se almacena para ser utilizado en sesiones sucesivas.

Estos tipos de personalización se aplican a dos tareas diferentes, filtrado (persistente) y recuperación (efímera). La personalización persistente se aplica utilizando IFTool [Asnicar *et al.* 97] como motor de filtrado. La personalización efímera se aplica para ayudar a los usuarios en la búsqueda de información. Esta ayuda puede ser de tipo terminológico, mediante el uso de un *tesaurus* para sugerir al usuario términos para expandir la consulta, y de tipo estratégico, mediante consejos consistentes en acciones a realizar por el usuario. En este último caso, el sistema monitoriza las acciones del usuario y utilizando una base de conocimiento es capaz de seleccionar sugerencias personalizadas basándose en la situación actual.

En Torii se muestran los documentos más relevantes ordenados por relevancia. La información mostrada sobre cada documento es: su título, sus autores, su relevancia en forma de barra continua, la fecha en la que fue recibido, y un enlace al documento completo.

2.8.22. SmartGuide

En Smart Guide [Gates *et al.* 98] el objetivo es suministrar a cada usuario un conjunto personalizado de las páginas web de la guía de referencia Web de la comunidad de usuarios del centro de investigación de supercomputadores de Mississippi (MCSR). La representación de los intereses del usuario se realiza en 3 pasos. El primer paso es responder a un pequeño formulario sobre cuestiones como intereses, experiencia, etc. De esta información se obtiene el modelo a largo plazo del usuario, el cual es representado en forma de estereotipos (estudiante, desarrollador Web, usuario de base de datos, etc.). Cada estereotipo está representado

por un vector de pesos de términos. El segundo paso es indicar al sistema, cada vez que se inicia una sesión, la necesidad de información actual en forma de consulta en lenguaje natural. Esta información constituye el modelo a corto plazo. Finalmente, durante la interacción con el sistema el usuario puede suministrar realimentación explícita sobre los elementos que recibe. El sistema combina los 3 tipos de intereses, estereotipos, consulta actual y realimentación para construir una única consulta obtenida a partir de una función polinomial de los vectores de pesos de términos que representan cada uno de los intereses.

Las fuentes de información que utiliza SmartGuide son fuentes heterogéneas distribuidas a lo largo de la red. Uno de los objetivos del sistema es precisamente la integración de tipos de información diferente en una interfaz única. Para ello representa cada fuente de información como objetos con una serie de atributos y valores asociados en forma de términos, además de almacenar qué peso se debe asociar a cada atributo en la indexación. La representación final de cada objeto es en forma de vector de pesos de términos.

La selección se realiza con la fórmula del coseno del MEV entre los vectores de pesos de términos que representan el modelo de usuario y cada una de las fuentes de información. El usuario puede realimentar el sistema, indicando su opinión sobre los elementos recibidos. Las posibilidades de realimentación son: bueno, malo y neutral. Los vectores de pesos de términos de los elementos votados como buenos y malos se tienen en cuenta en la siguiente selección de contenidos.

En Smart Guide se envían los títulos de las páginas Web seleccionadas para cada usuario ordenadas por relevancia con el modelo de usuario. Asociado a cada página aparecen tres botones de radio para indicar la realimentación (bueno, neutral, malo). Cada petición del usuario por una página Web tiene asociado un nombre de plantilla y parámetros opcionales. La plantilla corresponde a un fichero HTML mejorado que incluye una serie de huecos para situar objetos, así como información de como seleccionar y mostrar dichos objetos. La información de esta plantilla es procesada de tal forma que se muestra tal cual la parte que no corresponde a huecos para objetos. Sin embargo, para cada hueco se seleccionan los objetos que se deben incluir de acuerdo a los parámetros indicados (tipo de objeto, máximo número de objetos, cómo mostrar el objeto, etc.). En particular, hay una serie de parámetros que indican que mecanismo utilizar para la selección (correspondencia exacta con un identificador de objeto o mejor opción seleccionando a partir de una consulta), el umbral a partir del cual realizar la selección e instrucciones sobre como construir la consulta (sólo modelo de usuario, sólo contexto actual o combinación de ambos). El contexto actual está representado por el tipo de objeto que se está procesando y por un identificador de objeto. La apariencia final de cada objeto seleccionado es responsabilidad de otro módulo que encapsula la información sobre los objetos que maneja el sistema.

2.8.23. Fab

Fab [Balabanovic&Shoham97] es un sistema de recomendación de páginas Web, basado en filtrado colaborativo, donde los documentos se representan mediante los 100 términos con mayor peso $tf \cdot idf$. Cuando un usuario nuevo llega al sistema se le ofrece una serie de páginas aleatorias de entre un conjunto de páginas que son las que más le interesan al resto de usuarios que utiliza el sistema. De esta forma el usuario no empieza con un perfil vacío. Para ello se almacena la media de las evaluaciones de los usuarios en un perfil global.

En primer lugar se obtienen un conjunto de páginas interesantes para todos los usuarios, en base a una serie de temas, que posteriormente son enviadas a los usuarios para su poste-

rior selección. Para la obtención de las páginas se construyen una serie de agentes de búsqueda que tienen asociado un tema y un conjunto de palabras clave obtenidas a partir de las páginas evaluadas para ese tema. Por otro lado, existen una serie de agentes de selección, uno por usuario, que seleccionan las páginas más relevantes para los usuarios de acuerdo a su perfil. Cada cierto tiempo se envía a cada usuario una lista de nuevas páginas que debe evaluar en una escala de 7 puntos. Esas evaluaciones se utilizan para actualizar su agente de selección y también los agentes de búsqueda. Además, las páginas mejor evaluadas son enviadas a los agentes de selección de sus vecinos más cercanos.

2.8.24. P-Tango

P-Tango [Claypool *et al.* 99] es un sistema que permite la personalización del periódico electrónico Worcester Telegram and Gazzete Online. Este sistema utiliza una combinación de filtrado basado en el contenido y filtrado colaborativo. El modelo de usuario basado en contenido almacena las secciones del periódico en las que el usuario está interesado y un conjunto de palabras clave explícitas que el usuario puede especificar para cada sección. Además se almacenan un conjunto de palabras clave implícitas que se extraen de los documentos sobre los que el usuario ha emitido una valoración alta.

En este sistema se combinan técnicas de filtrado colaborativo y técnicas de filtrado de texto para personalizar un periódico electrónico. El filtrado basado en contenido mide la similitud entre palabras clave del modelo de usuario y de un documento como el número de palabras coincidentes por dos, dividido por el menor número de palabras de ambos elementos (modelo de usuario y documento). Cada parte del modelo aporta tres valores de similitud (secciones, palabras clave explícitas y palabras clave implícitas) con igual peso. Por otro lado, la predicción final se obtiene con una combinación ponderada de las similitudes obtenidas con el filtrado basado en el contenido y el filtrado colaborativo. Esta ponderación cambia por cada usuario y por cada artículo, para reflejar mejor los cambios en los gustos de los usuarios. Además las palabras clave implícitas se generan a partir de los documentos sobre los que el usuario realice valoraciones altas.

En la presentación de contenidos se muestra el título, sección, primeras frases y algunas veces el autor, de las noticias seleccionadas para el usuario, ordenadas por grado de interés. Al seleccionar un documento, aparece el texto completo del mismo. Además para realizar la realimentación se dispone de una barra coloreada continua que el usuario puede ajustar.

Se realiza una evaluación con 18 usuarios durante 3 semanas, 50 artículos por día, para comparar las técnicas de filtrado aisladas frente a su combinación. Se calcula el error absoluto medio entre las predicciones numéricas generadas por el sistema y las valoraciones emitidas por los usuarios. Durante la primera semana el filtrado de texto funciona mejor, durante la segunda los resultados son similares y durante la tercera es mejor el filtrado colaborativo. Por otro lado, durante las 3 semanas la combinación es algo mejor que cualquiera de los dos filtrados por separado.

2.8.25. GroupLens

En GroupLens[Resnick *et al.* 94; Konstan *et al.* 97] se combinan técnicas de filtrado colaborativo y técnicas de filtrado de texto para la personalización en grupos de noticias. En este caso el filtrado textual está basado en unos agentes de filtrado textual que evalúan los artículos según se van publicando. Estos agentes son considerados como si fueran usuarios que intro-

ducen muchos juicios y permiten disminuir el problema de la escasez de valoraciones. A partir de GroupLens surgió MovieLens (http://movielens.umn.edu) para la recomendación de películas [Good *et al.* 99; Sarwar *et al.* 01].

2.9. Resumen y conclusiones del capítulo

Las técnicas empleadas en los sistemas de personalización de contenidos revisados poseen cinco aspectos especialmente relevantes: la representación elegida para representar los documentos que se vayan a personalizar, la representación elegida para representar los modelos de los usuarios, la forma de seleccionar cuáles son los documentos más relevantes con respecto a un modelo de usuario, la forma de adaptar el modelo de usuario a través de la realimentación sobre los documentos recibidos y la forma de presentar los resultados obtenidos en el proceso de selección.

Los modelos para representar documentos más habituales son el booleano, el MEV, el probabilístico, y el basado en semántica latente. El MEV es el más utilizado y el más intuitivo.

Existen en la bibliografía distintas formas de definir los intereses de los usuarios que dependen básicamente de la cantidad y el tipo de información que se utilice para la representación de dichos intereses. Los más utilizados son los siguientes: categorías propias, términos, categorías, estereotipos, redes neuronales, redes semánticas, modelos a largo y corto plazo, y filtrado colaborativo. Los sistemas que combinan los intereses a corto y largo plazo del usuario permiten un modelado del usuario más completo, puesto que consideran distintos puntos de vista temporales en las necesidades de información de los usuarios.

Para realizar la selección de contenidos existen diversos algoritmos de clasificación dependiendo de las representaciones elegidas para el modelo de usuario y los documentos: fórmula del coseno del MEV, clasificador bayesiano ingenuo, vecino más cercano, reglas basadas en estereotipos, redes neuronales, redes semánticas y redes bayesianas.

Por otro lado, las técnicas de realimentación necesarias para poder conseguir un modelado dinámico del usuario se basan en la realimentación del usuario respecto de los elementos de información que se seleccionan según su perfil. La información obtenida se utiliza para actualizar el modelo de usuario de diversas formas según sea la representación elegida. Hay varias clases de técnicas de aprendizaje que se pueden utilizar para refinar los perfiles: aprendizaje directo, aprendizaje directo parcial, aprendizaje indirecto y comunidades de filtrado.

La técnica de filtrado colaborativo permite seleccionar los documentos en base a opiniones de usuarios con intereses similares, sin tener en cuenta el contenido de los propios documentos. Sin embargo, la utilización en solitario de un sistema de filtrado colaborativo puede ser poco efectiva debido al problema del primer evaluador, el problema de la escasez de valoraciones y el problema de la oveja negra. En particular, en sistemas donde la diseminación de la información se realiza a la vez para todos los usuarios no se puede utilizar esta técnica debido al problema del primer evaluador.

La mayoría de los sistemas de personalización utilizan opciones sencillas para la presentación de resultados: títulos, títulos y primeras frases, títulos y frases donde aparecen las palabras de la consulta resaltadas. En todo caso, en los sistemas de personalización examinados en la bibliografía es poco común la posibilidad de mostrar al usuario un resumen de los documentos seleccionados como interesantes y mucho menos que estos resúmenes se adapten a los intereses de los usuarios descritos en sus modelos de usuario.

La utilización de los resúmenes permite un ahorro de tiempo a los usuarios a la hora de detectar si un documento realmente le interesa sin tener que leerse el texto completo. Si además el resumen está personalizado según sus intereses, el usuario tardará aún menos tiempo, no sólo en decidir si le interesa o no, sino además en encontrar cual es la información que realmente le interesa de esa noticia.

Atendiendo al propósito del resumen se puede distinguir entre resúmenes indicativos, informativos, agregativos o críticos. Mientras que si tenemos en cuenta el enfoque los resúmenes pueden ser genéricos o adaptados al usuario.

Las técnicas que se utilizan para generar los resúmenes se agrupan en tres tipos: técnicas de extracción de frases, técnicas basadas en relaciones discursivas y técnicas de abstracción. Ante la variedad de tipos y dominios de los documentos disponibles, las técnicas de selección y extracción de frases resultan muy atractivas por su independencia del dominio y del idioma. El método se centra en una fase de análisis en el que se identifican los segmentos de texto (frases o párrafos, normalmente) que contienen la información más significativa. Durante esta fase se aplica un conjunto de heurísticas a cada una de las unidades de extracción. El grado de significación de cada una de ellas puede obtenerse mediante combinación lineal de los pesos resultantes de la aplicación de dichas heurísticas. Éstas pueden ser posicionales, lingüísticas, o estadísticas. El resumen resulta de concatenar dichos segmentos de texto en el orden en que aparecen en el documento original. La longitud del resumen puede variar según la tasa de compresión que se desee, en todo caso, un resumen con una tasa de compresión de entre el 20 y el 30% puede ser lo suficientemente informativo como para sustituir al documento original. Los posibles problemas de inconsistencia en el resumen resultante constituyen el principal inconveniente de esta aproximación, sin embargo, la mayoría de las frases incluidas en resúmenes manuales aparecen tal cual o con pequeñas modificaciones en el texto original.

Para poder medir la efectividad de los sistemas de personalización es necesario una serie de métodos que permitan medir hasta qué punto un sistema es bueno respecto a cada una de las tareas implicadas, esto es, selección, adaptación y presentación.

La mayoría de los criterios utilizados en la bibliografía para obtener los resultados de la evaluación son cuantitativos, es decir, asignan cantidades calculadas de diversas formas a las evaluaciones realizadas por el sistema en comparación con los juicios de relevancia determinados por los usuarios. Estos criterios proceden mayoritariamente del campo de la recuperación de información. Sin embargo, cada vez más se está optando por una evaluación cualitativa basada en opiniones de los usuarios, recogidas en cuestionarios, que muestran las impresiones de los usuarios sobre la utilización del sistema en diversos aspectos. En realidad, estas dos evaluaciones son complementarias y permiten visualizar el funcionamiento del sistema desde dos puntos de vista diferentes, desde el punto de vista del sistema y desde el punto de vista del usuario.

Para poder realizar una evaluación cuantitativa de un sistema un ingrediente fundamental son los juicios de relevancia que indican qué documentos son relevantes para cada usuario, de acuerdo a sus necesidades de información. En realidad estos juicios son una aproximación de la realidad, puesto que la relevancia de los documentos dependerá del usuario que esté efectuando la consulta y del contexto concreto en el que se realice la misma. De hecho, en los sistemas de personalización el conjunto de documentos relevantes es diferente para cada usuario, puesto que sus necesidades de información son diferentes. Por tanto, es necesario que cada usuario indique los juicios de relevancia sobre las noticias que recibe.

La forma de evaluar cuantitativamente una selección es comparar, de acuerdo a alguna métrica, los documentos seleccionados con los juicios de relevancia del usuario. Existen mu-

chas métricas, procedentes de la recuperación de información, que pueden ser agrupadas en varias categorías dependiendo del tipo de relevancia y del tipo de recuperación utilizadas [VanRijsbergen79; Salton&McGill83; Mizarro01]: relevancia y recuperación binarias, relevancia binaria y recuperación en forma de ranking, relevancia y recuperación como ranking, relevancia y recuperación como ranking de valores.

Cuando se comparan dos técnicas de selección o adaptación se debe poder afirmar que una técnica es mejor que otra de manera significativa. Para poder realizar esta afirmación hay que utilizar algún test de significancia estadística [Hull93]. La mayoría de los tests estándar basados en comparaciones entre parejas de valores producirán evidencia estadística si la probabilidad de que las diferencias entre los dos conjuntos de valores no se produzcan por azar es suficientemente pequeña, habitualmente menor que 0.05. Las parejas de valores que se utilizan para comparar son los resultados obtenidos para cada técnica según la métrica utilizada.

Para contrastar los resultados de las evaluaciones cuantitativas sobre los procesos de selección y adaptación es conveniente realizar una evaluación cualitativa, o evaluación centrada en el usuario, mediante el empleo de cuestionarios de evaluación que rellenen los distintos usuarios del sistema. Las preguntas de estos formularios sirven para ver la opinión de los usuarios sobre los distintas partes del sistema y son otra alternativa para medir la calidad del mismo.

En principio, la evaluación de la presentación de resultados se debería realizar utilizando una evaluación cualitativa donde los usuarios emitieran sus opiniones acerca de la calidad de esa presentación. Lógicamente esa calidad será mayor cuanto mejor ayuden al usuario a resolver sus necesidades de información. La mayoría de los comentarios realizados sobre la evaluación cualitativa para los procesos de selección y adaptación es aplicable a la presentación.

En todo caso, también se puede medir cuantitativamente la presentación de resultados teniendo en cuenta la utilización de esos resultados para resolver otra tarea de acceso a la información, o comparándolo con una presentación de resultados "ideal". En particular, cuando la presentación se realiza utilizando resúmenes lo que se intenta determinar cuantitativamente es cómo de adecuado o de útil es un resumen con respecto al texto completo [Hahn&Mani00]. Existen distintos enfoques que pueden clasificarse, según [Sparck-Jones&Galliers96], en directos (o intrínsecos) e indirectos (o extrínsecos). Los primeros se basan en el análisis directo del resumen a través de alguna medida que permita establecer su calidad. Los segundos basan sus juicios en función de su utilidad para realizar otra tarea.

Desde el comienzo de la investigación en generación de resúmenes automáticos hasta el momento han ido surgiendo distintas propuestas, individuales y colectivas, para intentar consensuar los distintos aspectos relacionados con su evaluación. Por ejemplo, la conferencia SUMMAC [Mani *et al.* 98] o las conferencias DUC [Harman&Marcu01; Hahn&Harman02; Radev&Teufel03]. Sin embargo, no se ha llegado todavía a un acuerdo pleno sobre cuál es el mejor método de evaluación debido a la complejidad de la tarea [Mani01].

El método de evaluación más utilizado para medir la relevancia del contenido de un resumen automático es la comparación con un resumen "ideal" construido manualmente. Sin embargo, la falta de concordancia entre jueces parece indicar que ese resumen no siempre existe. Por tanto, se puede dar el caso de un resumen contenga información relevante pero no se considere bueno porque no se parezca al resumen "ideal".

La evaluación indirecta sin jueces permite una evaluación mucho menos cara y mucho más extrapolable a grandes *corpora* de texto y a diversas técnicas de generación de resúmenes

con diversas tasas de compresión. Las evaluaciones se pueden realizar de manera más exhaustiva y automática.

Capítulo 3
PERSONALIZACIÓN DE CONTENIDOS WEB

3.1. Introducción

En este capítulo se va a describir el modelo de sistema de personalización de contenidos Web que se propone. El esquema fundamental en el que se basa la personalización es el de un sistema de filtrado de información basado en un modelo de usuario. En este caso la información que va a ser filtrada son documentos Web y lo que se va a ofrecer a los usuarios, el resultado del filtrado, es un único documento resultado con los contenidos adecuados según su modelo de usuario.

El modelo de sistema va a estar formado por los tres procesos que se describían en el capítulo anterior: selección de contenidos, adaptación del modelo de usuario y presentación de resultados. La utilización de técnicas de clasificación de texto combinada con la información contenida en los modelos de usuario va a ser el marco fundamental en el que se van a basar cada uno de los tres procesos. Estas técnicas serán aplicadas de distintas formas en las distintas versiones de los sistemas de personalización desarrollados en la tesis.

Aunque se tratará de describir las técnicas de personalización de contenidos Web de la manera más general posible, la aplicación de las mismas se realizará sobre el dominio concreto de los servicios de noticias de los periódicos electrónicos. Las principales características de este dominio han sido descritas en el apartado 1.3.

En cuanto al contenido del capítulo, éste se divide en 7 apartados que representan los distintos aspectos necesarios para realizar personalización de contenidos.

En el apartado 3.2 se indicará como se obtienen y representan los documentos Web a filtrar.

En el apartado 3.3 se describirá el modelado de usuario utilizado, describiendo cuál es la información de la que dispone el sistema para realizar la personalización. También se describirá cómo se produce la interacción entre dicho modelo y los usuarios. Además se indicará cómo se representan las distintas partes del modelo de usuario. En particular, se indicará como se obtiene el modelo a corto a plazo a través del proceso de adaptación del modelo de usuario.

En los apartados 3.4 y 3.5 se explicará cómo se realizan los procesos de selección y presentación de resultados, respectivamente. Todos ellos estarán basados en el modelo de usuario descrito en el apartado 3.3 y en la representación de la información indicada en el apartado 3.2.

En el apartado 3.6 se presentará un resumen de todos los parámetros que aparecen en el modelo del sistema.

Por último, en el apartado 3.7 se presentará un resumen y las conclusiones del capítulo.

3.2. Representación de los documentos Web

Los documentos se bajan de la web en forma de documentos HTML. Se extraen el título, categoría propia (sección en los periódicos), URL y texto de cada documento y se almacenan para ser procesados posteriormente. Para obtener una representación de las noticias se utiliza el modelo del espacio vectorial (MEV) [Salton&McGill83] aplicado al título y texto de las noticias. La categoría propia y el URL son metadatos asociados al documento que son utilizados en los distintos procesos de personalización de contenidos.

Los vectores de términos de los documentos son obtenidos eliminando las palabras vacías almacenadas en una lista de parada y almacenando las raíces de las palabras aplicando el extractor de raíces de Porter adaptado al español [Acero *et al.* 01]. Para obtener los pesos se utiliza la fórmula tf · idf, con tf como la frecuencia de apariciones de un término dentro de un documento e idf = $\log_2(N/df)$, siendo N el número de documentos de la colección.

Hay que tener en cuenta que en el dominio de los periódicos digitales, este proceso se repite cada día porque el conjunto de noticias es distinto cada día. Por lo tanto, los vectores de pesos de términos que representan a los documentos cambian cada día.

Se puede observar un ejemplo de noticia en el Apéndice II.

3.3. Modelado de usuario

La elección de un dominio u otro afecta a los aspectos que hay que potenciar en un modelo de usuario [Amato&Straccia99]. Aunque el dominio de los contenidos Web es muy amplio, se va a definir un modelo de usuario lo más genérico posible. Las únicas restricciones para que se pueda aplicar el modelo propuesto son que exista información textual asociada a los documentos Web y que exista una clasificación dependiente del dominio de aplicación.

Hay que tener en cuenta que cuando un usuario utiliza por primera vez un sistema de personalización puede no tener muy claro sus necesidades de información. En muchos sistemas se parte de un perfil vacío que se va actualizando por realimentación del usuario. Esto puede llegar a frustrar a los usuarios si el sistema inicialmente selecciona muchos documentos irrelevantes. Por tanto, es más interesante partir de un modelo de usuario inicial que puede ser obtenido a partir de las respuestas a un cuestionario o puede ser introducido de manera explícita por el usuario. En cualquier caso conviene que la forma de introducir ese perfil inicial sea lo más clara y sencilla posible para evitar que suponga una carga de trabajo excesiva.

Se propone un modelo o perfil de usuario navegable que representa los intereses de los usuarios desde diferentes puntos de vista. Este perfil almacena información sobre las preferencias del usuario relacionadas con el acceso a la información, es decir, qué información está buscando el usuario, con qué frecuencia desea recibir esta información, con qué formato y qué cantidad. El modelo de usuario almacena 3 tipos de información (ver Apéndice VI) [Díaz *et al.* 00a]:

- Información personal sobre el usuario: nombre, *login*, *password* y dirección de correo electrónico.

- Información sobre el formato de la información recibida: cuándo desea recibir información el usuario, qué días de la semana, y qué cantidad. La cantidad de información se especifica como un límite máximo en cuanto al número de documentos Web que se deben seleccionar.
- Información específica sobre los intereses del usuario según varios sistemas de referencia que serán los que se utilicen para realizar la personalización de los contenidos.

Establecer cuándo se desean recibir noticias y especificar un límite máximo en el número de documentos Web seleccionados evita la sobrecarga de los usuarios. Establecer un límite mínimo puede llegar a ser contraproducente porque podría llevar a la inclusión de información no relevante para el usuario.

El modelo de usuario está basado en varios sistemas de referencia para obtener distintas vistas o descripciones de los intereses de un usuario. Los procesos de personalización combinarán la información de los sistemas de referencia para realizar su trabajo [Díaz01].

El modelo de usuario propuesto consiste en la combinación de dos tipos de intereses de un usuario: a largo y a corto plazo (Figura 3.1).

Cuando un usuario utiliza un sistema de filtrado de información define unos intereses más o menos estáticos que se almacenan en su modelo de usuario. En el caso de la personalización Web también aparece esta situación en la cual el usuario tiene unos intereses que permanecen fijos a lo largo del tiempo, ya sea por su trabajo, sus aficiones, etc., estos intereses son los que estarán modelados por los intereses a largo de plazo del modelo de usuario. Sin embargo, las necesidades de los usuarios varían con el tiempo, sobre todo como efecto directo de su interacción con la información [Belkin97; Billsus&Pazzani00]. Por tanto, es bastante probable que los intereses de los usuarios no permanezcan estáticos sino que vayan variando según van recibiendo información. Estas variaciones de interés a lo largo del tiempo estarán modelados por los intereses a corto plazo.

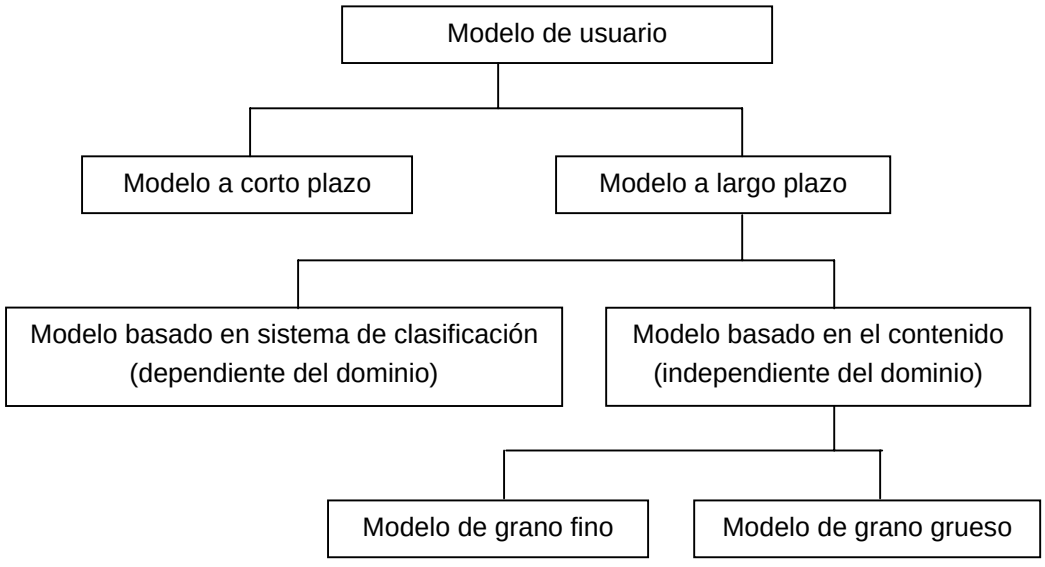


Figura 3.1. Modelo de usuario.

3.3.1. Modelo a largo plazo

Para modelar los intereses a largo plazo se utilizan dos marcos de referencia: uno basado en un sistema de clasificación dependiente del dominio y otro basado en el contenido de los documentos (independiente del dominio) (Figura 3.1).

Una de las alternativas para representar explícitamente el interés del usuario es mediante un conjunto de categorías propio del dominio utilizado. Un usuario debe seleccionar aquellas categorías a las cuales pertenecen los documentos que él considera interesantes. Estos conjuntos de categorías propias suelen estar prefijadas por el suministrador de los documentos Web. Dominios diferentes requieren diferentes sistemas de categorías, o al menos, diferentes sistemas de categorías son considerados más efectivos para dominios específicos. Por ejemplo, el sistema de categorías de los periódicos lo constituyen las secciones (nacional, internacional, economía, sociedad, cultura, etc.) y el una tienda son los géneros de los productos que vende (música pop, música clásica, literatura, viajes y ocio, etc.) [Díaz&Gervás00].

Para cada categoría propia el usuario asigna un peso que indica su importancia. Estos pesos para cada categoría para cada usuario se representan en una matriz C_{cu}, donde cada fila representa una categoría propia c y cada columna un usuario u. Los valores numéricos asociados a los pesos se extraen a partir de una asignación del usuario basada en 4 posibles valores: no me interesa, me interesa algo, me interesa, me interesa mucho (ver Apéndice VI). Estos valores corresponden a los valores numéricos 0, 0.33, 0.66, 1.

El sistema de referencia basado en el contenido de los documentos se subdivide a su vez en dos sistemas de referencia, un modelo de grano fino basado en palabras clave y un modelo de grano grueso basado en categorías generales independientes del dominio (Figura 3.1).

Para construir el modelo de grano fino el usuario puede introducir un conjunto de palabras clave para definir sus intereses. El usuario debe asignar un peso a cada una estas palabras para indicar la importancia relativa de cada una de ellas. La aparición de estas palabras en los documentos indica que éstos pueden interesar el usuario. Las palabras clave se almacenan, para cada usuario u, como un vector de pesos de términos (p_u). Los pesos se obtienen a partir de las valoraciones introducidas por el usuario de la misma manera que con las categorías propias.

Otra alternativa para representar explícitamente el interés del usuario es mediante un conjunto de categorías generales independientes del dominio utilizado (modelo de grano grueso). El usuario debe seleccionar las categorías generales en las que está interesado, así como su grado de interés en cada una de ellas. El sistema de categorías generales utilizado es el que corresponde a las categorías de primer nivel de Yahoo! España. Se ha elegido este sistema de clasificación por la familiaridad de los usuarios con los directorios de búsqueda de Yahoo!. Son las siguientes 14 categorías: Arte y cultura, Ciencia y tecnología, Ciencias sociales, Deportes y ocio, Economía y negocios, Educación y formación, Espectáculos y diversión, Internet y ordenadores, Materiales de consulta, Medios de comunicación, Política y gobierno, Salud, Sociedad y Zonas geográficas. Un ejemplo de página de Yahoo! se muestra en el Apéndice IV.

La representación de las categorías generales se obtiene a partir de un conjunto de entrenamiento que almacena un conjunto de documentos que pertenecen a las distintas categorías que se van a utilizar. De esta manera, se puede obtener para cada categoría i una representación formada por un vector de pesos de términos (g_i).

Inicialmente se ha obtenido la representación de cada categoría como vectores de pesos de términos obtenidos a partir de la siguiente información: el nombre de la categoría de primer nivel de Yahoo!, el nombre de las subcategorías de segundo nivel y la descripción de las

páginas que aparecen en la página web asociada a la categoría de primer nivel. Toda esta información aparece en la página de la categoría de primer nivel y es pre-procesada de la misma manera que los documentos Web, es decir, aplicando filtrado de palabras de vacías y extractor de raíces. Por tanto el conjunto de entrenamiento es simplemente un documento por categoría y el método utilizado es el algoritmo de Rocchio teniendo sólo en cuenta el entrenamiento positivo. Esta representación se utilizó en los experimentos preliminares.

Sin embargo, esta representación resultó ser escasa para algunas categorías por lo que se amplió utilizando para cada categoría las páginas de las categorías de segundo nivel, además de la página de la categoría de primer nivel. Esta representación se utilizó para la última versión de sistema de personalización.

Por otro lado, los pesos asignados por el usuario a cada una de las categorías generales se representan en una matriz G_{gu}, donde cada fila representa una categoría general g y cada columna un usuario u. Los pesos se obtienen a partir de las valoraciones introducidas por el usuario de la misma manera que con las categorías propias.

El usuario puede editar y modificar su perfil a largo plazo, dando pesos a los distintos elementos de cada uno de los sistemas de referencia. También puede añadir, modificar y borrar sus palabras clave. En el Apéndice VI se puede observar un ejemplo de utilización de la interfaz gráfica para la definición del perfil de un usuario.

3.3.2. Modelo a corto plazo

El modelo a corto plazo almacena los intereses del usuario que varían a lo largo del tiempo. La representación del modelo a corto plazo para el usuario u se realiza a través de un vector de pesos de términos (t_u) que se obtiene a partir del proceso de adaptación del modelo de usuario.

3.3.2.1. Adaptación del modelo de usuario

La adaptación del modelo de usuario es necesaria por que las necesidades de los usuarios varían con el tiempo, principalmente como efecto de la interacción con la información que reciben [Belkin97; Billsus&Pazzani00]. Por esta razón el modelo de usuario debe ser capaz de adaptarse a esos cambios de interés. El proceso de adaptación del modelo de usuario permite la obtención/actualización del modelo a corto plazo a partir de la realimentación del usuario.

El punto de partida del proceso de adaptación es la presentación personalizada (ver apartado 3.5) de cada uno de los M^+ (ver apartado 3.6) documentos Web que el sistema ha encontrado más relevantes según el modelo del usuario. El usuario puede interactuar con el sistema mediante la realimentación sobre cada uno de los elementos de información recibidos. Esta realimentación puede ser positiva, negativa o nula. En el primer caso, indica que el usuario está interesado en el documento y desea seguir recibiendo documentos que traten sobre el mismo tema. En el segundo, lo que indica es que el usuario no está interesado y no desea seguir recibiendo documentos del mismo tema. El tercero caso implica que el usuario opta por no realizar realimentación sobre un determinado documento, esto indica que le resulta indiferente.

[+] Las letras latinas mayúsculas corresponden a parámetros del sistema cuyos valores serán fijados para realizar los experimentos (ver apartado 3.6)

Los términos extraídos de los documentos sobre los que el usuario ha realizado algún tipo de realimentación (positiva o negativa) sirven para obtener los términos de realimentación (t_u) que constituyen el modelo a corto plazo o del usuario u. Estos términos tienen un peso asociado que depende de varios factores que se indican más adelante. Los términos de realimentación se añaden al modelo tal cual, si no existían previamente, o actualizan el valor de su peso, si ya estaban presentes en el modelo.

Como este marco de referencia está pensado para modelar los intereses a corto plazo de un usuario se utiliza un algoritmo para decrementar el peso de los términos del modelo con el paso del tiempo. Cada día se obtienen los nuevos pesos de los términos restando una cierta cantidad D^+ (ver apartado 3.6) de los pesos del día anterior. Si el peso de alguno de los términos de realimentación llega a tener un valor menor o igual que cero, entonces este término es eliminado del modelo.

La actualización de los términos de realimentación para un determinado usuario consiste en la sustracción de la cantidad D de los valores de todos sus términos de realimentación y de la eliminación de aquellos que alcancen un valor menor o igual que cero. Esta operación se realiza todos los días antes de que la nueva información sobre realimentación sea tenida en cuenta. Los valores resultantes de esta operación, para cada usuario u y para cada término de realimentación t, serán los valores iniciales de interés de cada término t para cada usuario u y se denominarán I_{tu}.

A continuación se aplica una adaptación del algoritmo de [Nakashima&Nakamura97] (ver apartado 2.5.1) para obtener los términos de realimentación, esto es, el proceso de realimentación de los términos de su modelo "consciente" es utilizado para obtener el modelo a corto plazo en nuestra propuesta. Además, el conjunto $R_u(-)$ son los documentos sobre los que el usuario realimenta negativamente. Para mayor claridad del proceso se va a exponer a continuación la adaptación del algoritmo utilizada.

Los siguientes cálculos deben realizarse cada día para cada usuario u para obtener los nuevos términos de realimentación, ya que los documentos de cada día son diferentes y la realimentación para un determinado día es diferente para cada usuario.

En primer lugar hay que tener en cuenta que sólo se consideran los M primeros documentos en el ranking del usuario para obtener los términos de realimentación. La justificación es que, aunque se pueda disponer de los juicios de realimentación sobre todas las noticias, el usuario en un comportamiento normal del sistema sólo va a recibir las M primeras y por ello, los efectos de la realimentación se deben realizar teniendo sólo en cuenta estas noticias.

Se dividen los documentos realimentados por cada usuario u, entre los M primeros, en dos conjuntos: $R_u(+)$, conjunto de documentos con realimentación positiva y $R_u(-)$, conjunto de documentos con realimentación negativa. Un documento puede pertenecer a uno de los dos conjuntos o a ninguno de los dos, pero no puede pertenecer a los dos a la vez. El conjunto de todos los documentos sobre los que el usuario u ha efectuado algún tipo de realimentación, positiva o negativa, se denomina R_u. La obtención de los términos t_u que representan la realimentación para un usuario u se realiza a través de los pasos que se describen a continuación.

Para la selección/actualización de los nuevos términos de realimentación se preprocesan primero todos los documentos de la misma forma a como se realiza en el proceso de representación de los documentos, esto es, se filtran todas las palabras vacías con una lista de parada y después se aplica el extractor de raíces adaptado al español. Es decir, que el punto de partida del proceso de adaptación son los términos de la representación de los documentos, con su frecuencia asociada (tf).

Se define el valor de acceso a_{tdu} para el término t en el documento d para el usuario u de la siguiente manera:

$$a_{tdu} = \begin{cases} P(T \cdot titulo_{td} + cuerpo_{td}) & si\ d \in R_u(+) \\ -N(T \cdot titulo_{td} + cuerpo_{td}) & si\ d \in R_u(-) \end{cases} \quad (3.1)$$

donde:

- $titulo_{td}$ es la frecuencia del término t en el título del documento d.
- $cuerpo_{td}$ es la frecuencia del término t en el cuerpo del documento d.
- P^+ es el peso para la realimentación positiva.
- N^+ es el peso para la realimentación negativa.
- T^+ es el peso del título.

De esta manera, un término tendrá valor de acceso alto si aparece con mucha frecuencia en títulos y en cuerpos de documentos con realimentación positiva, y tendrá valor bajo si aparece en documentos con realimentación negativa. La utilización de los términos que aparecen en el título está justificada por el hecho de que la realimentación de los usuarios se basa mucho más en el título que en el contenido de los documentos [Díaz *et al.* 05b]. Este valor de acceso mide la representatividad de los términos que aparecen en los documentos realimentados.

El porcentaje de actualización p_{tu} de un término t para un usuario u se obtiene normalizando los resultados anteriores a través de todos los documentos:

$$p_{tu} = \frac{\sum_{d \in R_u} a_{tdu}}{\max(\left|\sum_{d \in R_u} a_{tdu}\right|)} \quad (3.2)$$

De esta manera, se suman los valores de acceso de todos los términos y se normalizan de tal forma que el término con mayor porcentaje de actualización tenga valor 1, y el resto tengan valores entre 0 y 1.

El nuevo valor de interés N_{tu} para el término t para el usuario u se obtiene según la siguiente fórmula:

$$N_{tu} = \begin{cases} I_{tu} + (V \cdot (1 - I_{tu}) \cdot p_{tu}) & si\ p_{tu} \geq 0 \\ I_{tu} - (V \cdot I_{tu} \cdot \left|p_{tu}\right|) & si\ p_{tu} < 0 \end{cases} \quad (3.3)$$

donde:

- V^+, varía entre 0 y 1 e indica la velocidad de cambio del valor de interés de un término, es decir, cuanto mayor sea esta velocidad más deprisa cambiará el valor de interés, en el sentido de que habrá más diferencia entre el valor inicial y el nuevo valor.
- I_{tu} es el interés inicial del término t para el usuario u.

El resultado final de este proceso es un conjunto de términos por cada usuario ordenados según su nuevo valor de interés. Se selecciona un subconjunto de ellos, los R^+ más relevantes, para formar parte de los términos de realimentación (t_u) del modelo a corto plazo del

[+] Las letras latinas mayúsculas corresponden a parámetros del sistema cuyos valores serán fijados para realizar los experimentos (ver apartado 3.6)

usuario u. De estos términos seleccionados sólo se añadirán aquellos que sean diferentes de los términos de realimentación ya existentes, el resto actualizarán su peso asociado.

3.3.3. Combinación de modelos a largo y corto plazo

Una contribución importante de esta tesis consiste en probar cuál es la mejor forma de combinar los distintos sistemas de referencia para obtener la mejor solución posible. Para lograr este propósito, cada uno de los 4 sistemas de referencia descritos (categorías propias, categorías generales, palabras clave y términos de realimentación) tiene un peso que indica su importancia en la obtención de los resultados finales. De esta forma, cada una de las 4 dimensiones puede ser definida y controlada durante los experimentos. Esto permite separar los intereses a largo y corto plazo, y separar las categorías propias, las categorías generales y las palabras clave, dentro del modelo a largo plazo.

Se puede observar un ejemplo de representación interna de un modelo de usuario completo en el Apéndice II.

3.4. Selección de contenidos

La selección de contenidos se refiere a la elección entre todos los documentos de entrada de aquellos que son más relevantes para un usuario dado, según su perfil o modelo. Una vez fijadas la representación de los documentos y la representación del modelo de usuario se puede calcular la similitud entre ambas representaciones.

A partir de ahora se hablará indistintamente de similitud entre documentos y modelo de usuario o relevancia de los documentos para el modelo de usuario, teniendo en cuenta que este tipo de relevancia se refiere a la relevancia calculada por el sistema y no a la relevancia indicada por los usuarios.

Puesto que existen distintos sistemas de referencia en el modelo de usuario se va a indicar cómo se obtiene la selección de contenidos con respecto a cada uno de ellos y posteriormente se explorarán las distintas combinaciones. Para establecer las combinaciones se tendrá en cuenta la relevancia obtenida con cada sistema de referencia y el peso relativo utilizado en cada una de las combinaciones. En cualquier caso el resultado final será un ranking de los documentos ordenado por similitud con el modelo de usuario. Los documentos más similares (más relevantes), es decir, la parte superior del ranking, se seleccionan para cada usuario, respetando el límite máximo indicado en su modelo (M).

3.4.1. Selección con respecto al modelo a largo plazo

La selección de contenidos respecto al modelo a largo plazo tiene contribuciones provenientes de los distintos sistemas de referencia representados en el modelo, es decir, categorías propias, categorías generales y palabras clave.

3.4.1.1. Selección con respecto a las categorías propias

Como cada documento Web tiene una categoría pre-asignada, la selección respecto este marco de referencia es inmediata. Todos los documentos son procesados para comprobar si

pertenecen a alguna de las categorías seleccionadas por el usuario. La relevancia[+] r_{du}^{c} entre un documento d, que pertenece a una categoría propia c, y un modelo de usuario u es directamente el valor C_{cu} asignado a la categoría propia c por el usuario u.

$$r_{du}^{c} = C_{cu} \tag{3.4}$$

A partir de la ecuación (3.4) se obtiene un ranking de documentos con respecto a las categorías propias c de un modelo de usuario dado u. Todos los documentos que pertenezcan a la misma categoría propia tendrán la misma relevancia y aparecerán en posiciones consecutivas del ranking.

3.4.1.2. Selección con respecto a las palabras clave

La relevancia r_{du}^{p} entre un documento d y las palabras clave p de un modelo de usuario u es calculada mediante la fórmula de similitud del coseno entre la representación del documento y la representación de las palabras clave:

$$r_{du}^{p} = sim(d_d, p_u) \tag{3.5}$$

donde:

- p_u es el vector de pesos de términos de las palabras clave del usuario u.
- d_d es el vector de pesos de términos del documento d.
- sim es la fórmula del coseno del MEV.

A partir de la ecuación (3.5) se obtiene un ranking de documentos con respecto a las palabras clave p de un modelo de usuario u.

3.4.1.3. Selección con respecto a las categorías

La relevancia r_{dg_i} entre un documento d y una categoría general g_i, se calcula mediante la fórmula del coseno entre la representación del documento y la representación de la categoría:

$$r_{dg_i} = sim(d_d, g_i) \tag{3.6}$$

donde:

- g_i es el vector de pesos de términos de la categoría general i.
- d_d es el vector de pesos de términos del documento d.
- sim es la fórmula del coseno del MEV.

A partir de la ecuación (3.6) se obtiene un ranking de documentos con respecto a la categoría general i. Esta fórmula ha de ser aplicada tantas veces como categorías generales haya, en nuestro caso, 14. Para cada aplicación se obtiene el ranking de documentos asociado a la categoría general correspondiente.

[+] Los superíndices muestran la parte del modelo a la que se refiere la relevancia, esto es, superíndice c en la primera relevancia indica que este es el valor para las categorías propias, g, para las categorías generales, p, para palabras clave y t, para términos de realimentación, también, l, para largo plazo, y o, para corto plazo.

Esta relevancia es independiente de los modelos de usuario, puesto que la representación de las categorías generales sólo depende de la colección de entrenamiento y no de los modelos de usuario.

La relevancia r_{du}^{g} entre un documento d y las categorías generales g almacenadas en el modelo de usuario u es calculada mediante la siguiente fórmula:

$$r_{du}^{g} = \frac{\sum_{i=1}^{14} G_{iu} r_{dg_i}}{\sum_{i=1}^{14} G_{iu}} \tag{3.7}$$

donde G_{iu} muestra la importancia asignada a la categoría general i por el usuario u.

A partir de la ecuación (3.7) se obtiene un ranking de documentos con respecto a las categorías generales g de un modelo de usuario dado u.

3.4.1.4. Selección combinando los tres sistemas de referencia

Cuando todos los documentos han sido ordenados con respecto a las diferentes fuentes de relevancia, los resultados son integrados utilizando la combinación particular que es asignada a cada uno de los sistemas de referencia. Por tanto, la relevancia total r_{du}^{l} entre un documento d y el modelo a largo plazo l de un modelo de usuario u se calcula con la siguiente fórmula:

$$r_{du}^{l} = \frac{\alpha r_{du}^{c} + \beta r_{du}^{g} + \chi r_{du}^{p}}{\alpha + \beta + \chi} \tag{3.8}$$

donde las letras griegas α, β, y χ muestran la importancia asignada a cada uno de los sistemas de referencia (α, para categorías propias, β, para categorías generales, χ, para palabras clave). Para que esta combinación sea significativa, la relevancia obtenida a partir de cada sistema de referencia debe ser normalizada con respecto a los mejores resultados para la colección de documentos que se esté utilizando. Por ejemplo, si la mayor relevancia asignada a los documentos de un determinado día en términos de palabras clave, para un usuario, es 0.42, todas las relevancias obtenidas en términos de palabras clave para ese mismo día, para ese usuario, deben ser normalizadas antes de combinarlas con las relevancias obtenidas para los otros sistemas de referencia. Intuitivamente, esto corresponde a ajustar las escalas para que el documento que obtenga mayor relevancia para cada sistema obtenga relevancia igual a 1. Esto asegura que las distintas relevancias provenientes de los distintos sistemas puedan ser mezcladas sin que los valores absolutos de relevancia tengan que ser de la misma magnitud. De hecho, la relevancia obtenida a partir de las categorías propias es 0 ó 1 (un documento pertenece o no pertenece a una categoría propia), mientras que, por ejemplo, el sistema de referencia de palabras clave puede asignar un valor de relevancia bajo a un documento que contenga dos palabras clave si el resto de las palabras clave no están contenidos en el mismo. Sin embargo, si esa es la mejor elección disponible en función de las palabras clave, es necesario que ese documento con dos palabras clave tenga una relevancia que pueda competir con aquellos documentos que pertenezcan a las categorías propias elegidas por el usuario, por tanto su relevancia es normalizada a 1 antes de que las relevancias de ambos sistemas sean combinadas.

El ratio entre β e χ representa la importancia relativa asignada, dentro del modelo basado en contenido del modelo a largo plazo, a la clasificación basada en las categorías generales con respecto a la clasificación basada en palabras clave.

El ratio entre α y β+χ representa la importancia relativa asignada dentro del modelo a largo plazo a la clasificación dependiente del dominio (categorías propias) con respecto a la clasificación independiente del dominio (categorías generales y palabras clave). α, β, χ actuarán como variables de control en los experimentos.

A partir de la ecuación (3.8) se obtiene un ranking de documentos con respecto al modelo a largo plazo l de un usuario u.

3.4.2. Selección con respecto al modelo a corto plazo

La relevancia r_{du}^{o} entre un documento d y el modelo a corto plazo o de un modelo de usuario u es calculada de manera similar a la calculada con respecto a las palabras clave pero utilizando el vector de pesos de términos de realimentación t, del modelo de usuario u, obtenido en el proceso de adaptación del modelo de usuario:

$$r_{du}^{o} = r_{du}^{t} = sim(d_d, t_u) \tag{3.9}$$

donde:

- t_u es el vector de pesos de términos de los términos de realimentación del usuario u.
- d_d es el vector de pesos de términos del documento d.
- sim es la fórmula del coseno del MEV.

A partir de la ecuación (3.9) se obtiene un ranking de documentos con respecto al modelo a corto plazo o de un usuario u.

3.4.3. Selección con respecto a la combinación de largo y corto plazo

Cuando todos los documentos han sido ordenados con respecto a las diferentes fuentes de relevancia, los resultados son integrados utilizando la combinación particular que es asignada a cada uno de los sistema de referencia. Por tanto, la relevancia total r_{du} entre un documento d y un modelo de usuario u se calcula con la siguiente fórmula:

$$r_{du} = \frac{\alpha r_{du}^{c} + \beta r_{du}^{g} + \chi r_{du}^{p} + \varepsilon r_{du}^{o}}{\alpha + \beta + \chi + \varepsilon} \tag{3.10}$$

donde las letras griegas α, β, χ y ε muestran la importancia asignada a cada uno de los sistemas de referencia (α, para categorías propias, β, para categorías, χ, para palabras clave, ε, para el modelo a corto plazo). Para que esta combinación sea significativa, la relevancia obtenida a partir de cada sistema de referencia debe ser normalizada con respecto a los mejores resultados para la colección de documentos que se esté utilizando.

El valor relativo de ε con respecto a $\alpha+\beta+\chi$ representa la importancia relativa asignada al modelo a corto plazo con respecto al modelo a largo plazo. α, β, χ y ε actuarán como variables de control en los experimentos.

A partir de la ecuación (3.10) se obtiene un ranking de documentos con respecto a un modelo de usuario u. Los primeros M documentos del ranking son seleccionados para cada usuario.

3.5. Presentación de resultados

La presentación de resultados consiste en, una vez seleccionados los documentos que le interesan a un usuario, generar un único contenido que contenga estos documentos como elementos de información. La generación de este contenido se personaliza teniendo en cuenta las distintas preferencias indicadas por el usuario en su perfil. Los contenidos generados se plasman en un documento web en formato HTML que puede ser enviado por correo electrónico o visualizado como una página web.

El formato del documento web generado es el siguiente [Díaz *et al.* 00]:

- Un título con la fecha y el nombre del usuario.
- Un enlace al modelo de usuario para permitir la edición del mismo.
- Una breve descripción de los intereses del usuario (tal y como aparecen en su perfil).
- Los documentos seleccionados, presentados de mayor a menor relevancia y respetando el límite máximo definido por el usuario (M).
- Para cada documento seleccionado para el usuario se presenta la siguiente información:
    - Título.
    - Nombre del autor.
    - Nombre de la categoría propia a la que pertenece.
    - Fuente.
    - Relevancia.
    - Resumen.
    - Enlace al documento completo en la web del dominio.
    - Iconos de realimentación (negativa/positiva), donde el usuario debe pulsar para realizar la realimentación.

En la Figura 3.2 se puede observar un ejemplo de mensaje enviado por el segundo sistema de personalización.

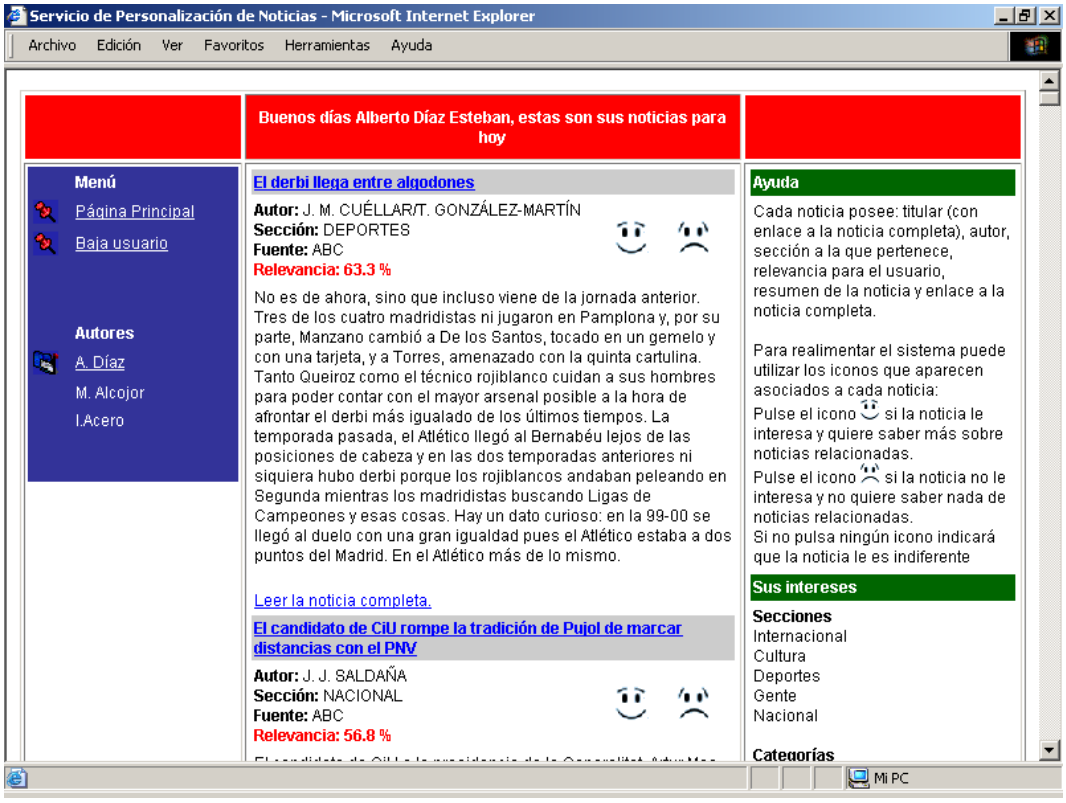


Figura 3.2. Ejemplo de mensaje enviado por el segundo sistema de personalización.

El comportamiento habitual de los sistemas de personalización es presentar a los usuarios el título y las primeras líneas de los elementos de información relevantes. Esta información suele ser insuficiente para que el usuario detecte si el elemento es realmente relevante para él. Esto hace que el usuario tenga que inspeccionar el texto completo, lo cual supone un coste adicional en tiempo. Además, puede ocurrir que la estructura y el tamaño del documento no sean adecuados y el usuario tenga dificultades para encontrar la información en la que está interesado y poder decidir sobre la relevancia del documento.

La utilización de los resúmenes permite un ahorro de tiempo a los usuarios a la hora de detectar si un documento realmente le interesa sin tener que leerse el texto completo. Si además el resumen está personalizado según sus intereses, el usuario tardará aún menos tiempo no sólo en decidir si le interesa o no, sino además en encontrar cuál es la información que realmente le interesa de esa noticia.

El resumen de cada documento es extraído automáticamente del texto completo de cada documento aplicando técnicas de selección y extracción de frases, donde una frase o sentencia es el conjunto de palabras entre dos puntos que indiquen final de frase.

Las técnicas de selección y extracción de frases resultan muy atractivas por su independencia del dominio y del idioma. Además estas técnicas pueden tener en cuenta la información almacenada en el modelo de usuario permitiendo personalizar los resúmenes, es decir, generar resúmenes adaptados al perfil del usuario. Esta es precisamente la principal aportación en la personalización de contenidos, la generación de resúmenes personalizados: un

resumen puede no ser útil para un usuario sino se seleccionan las frases más relacionadas con sus intereses.

A continuación se van a describir las distintas técnicas utilizadas, así como sus posibles combinaciones. Se utilizan tres heurísticas de selección de frases o sentencias para construir los resúmenes. Las dos primeras se utilizan para construir resúmenes genéricos, mientras que la tercera sirve para generar resúmenes personalizados. Se indicarán las posibles combinaciones para construir los resúmenes.

Las tres heurísticas tienen un objetivo común: asignar a cada sentencia un valor que indique su relevancia para formar parte del resumen. La selección final consistirá en las L%[+] sentencias que tengan mayor valor de relevancia. Estas sentencias se concatenarán respetando su orden original para evitar inconsistencias.

Se va a describir cada heurística y se comentará como se pueden modificar los distintos parámetros para obtener resúmenes genéricos o personalizados.

3.5.1. Resúmenes genéricos

En primer lugar, se van a describir 2 heurísticas que son independientes del usuario al que vaya dirigido el resumen. La primera tiene en cuenta la posición de cada frase en el documento y la segunda tiene en cuenta las palabras significativas que aparecen en cada frase.

3.5.1.1. Heurística de posición

En muchos dominios (p. ej.: periódicos digitales), las primeras sentencias de un texto suelen constituir un buen resumen del contenido del texto completo. Esta heurística asigna los valores más altos a las 5 primeras sentencias de un texto [Edmunson69]. En nuestro trabajo los valores específicos elegidos para estas 5 sentencias se muestran en la Tabla 3.1. Las sentencias de la sexta en adelante se les asigna el valor 0. Estos valores generan los pesos A_{sd} para cada sentencia s dentro de cada documento d. Estos valores son independientes del usuario u que se esté considerando.

[+] Las letras latinas mayúsculas corresponden a parámetros del sistema cuyos valores serán fijados para realizar los experimentos (ver apartado 3.6)

Número de sentencia	Valor asignado
1	1.00
2	0.99
3	0.98
4	0.95
5	0.90
Resto	0.00

Tabla 3.1. Valores asignados a las sentencias de un documento según la heurística de posición.

A partir de esta heurística se obtiene un ranking de sentencias para cada documento con respecto a la posición que éstas ocupen en el documento, es decir, que el ranking será el orden de las frases en el documento.

3.5.1.2. Heurística de palabras significativas

Cada texto tiene un conjunto de palabras significativas o "temáticas" que son representativas de su contenido. Esta heurística extrae las S^+ palabras no vacías más significativas para cada texto y calcula cuántas de estas palabras aparecen en cada sentencia. Las frases tendrán mayor valor cuanta mayor densidad de palabras significativas contengan [Kupiec *et al.* 95; Teufel&Moens97].

Para obtener las S palabras más significativas de cada documento se utiliza el peso obtenido por cada palabra en la representación de los documentos, esto es, el peso tf · idf obtenido por cada palabra en la indexación de los documentos.

Para obtener el valor asignado B_{sd} a cada sentencia s dentro del documento d usando la heurística de palabras significativas, el número de palabras significativas que aparecen en una sentencia se divide por el número total de palabras no vacías que aparecen en la sentencia. El objetivo de este cálculo es dar más peso a aquellas sentencias que tengan una mayor densidad de palabras significativas [Teufel&Moens97]. Estos valores son independientes del usuario u que se esté considerando.

A partir de esta heurística se obtiene un ranking de sentencias para cada documento con respecto a las palabras significativas que contiene cada frase.

3.5.1.3. Combinación de las heurísticas

Para obtener el valor asociado G_{sd} entre cada sentencia s dentro de un documento d utilizando la combinación de las dos heurísticas anteriores se aplica la siguiente ecuación:

$$G_{sd} = \frac{\varphi A_{sd} + \gamma B_{sd}}{\varphi + \gamma} \tag{3.11}$$

Los parámetros φ e γ permiten el ajuste entre las diferentes heurísticas, dependiendo de si se prefiere utilizar la heurística de posición (φ) o la de palabras significativas (γ). Para que esta

[+] Las letras latinas mayúsculas corresponden a parámetros del sistema cuyos valores serán fijados para realizar los experimentos (ver apartado 3.6)

combinación sea significativa, la relevancia obtenida a partir de cada sistema de referencia debe ser normalizada con respecto a los mejores resultados para la colección de documentos que se esté utilizando.

El valor relativo entre φ e γ representa la importancia relativa asignada a las heurística de posición con respecto a la heurística de palabras clave. Los valores de φ e γ actuarán como variables de control en los experimentos.

A partir de la ecuación (3.11) se obtiene un ranking de sentencias para cada documento con respecto a las heurísticas genéricas, es decir, independientes del usuario.

3.5.2. Resúmenes personalizados

A continuación se va a describir la heurística de personalización, es decir, la que utiliza información dependiente del usuario al que vaya dirigido el resumen.

3.5.2.1. Heurística de personalización

El objetivo de esta heurística es seleccionar aquellas frases que son más relevantes para un modelo de usuario determinado. El potencial de la personalización de los resúmenes es grande, porque un documento que se resumiera de manera genérica podría no ser útil para un usuario, mientras que si seleccionaran las frases más similares a su modelo de usuario podría convertirse en un resumen mucho más interesante al contener la información que realmente le interesa al usuario.

El cálculo de los valores P_{sdu} asociados a cada sentencia s dentro de un documento d para un usuario u, se obtendrán calculando la similitud entre el modelo de usuario y cada una de las sentencias del documento a resumir.

A partir esta heurística se obtiene un ranking de sentencias para cada documento con respecto a la heurística de personalización.

A continuación se van a presentar las distintas posibilidades a la hora de utilizar esta heurística. Dependiendo de la parte del modelo de usuario que utilicemos obtendremos una personalización u otra.

3.5.2.2. Personalización utilizando las palabras clave

Para obtener resúmenes personalizados utilizando las palabras clave del modelo a largo plazo l de un usuario u hay que seguir el siguiente proceso. Primero, se obtiene el vector de pesos de términos p_u asociado a las palabras clave del modelo a largo plazo l del usuario u. Después, se obtienen los vectores de pesos de términos asociados a cada una de las sentencias del documento a resumir. Finalmente, el valor P_{sdu}^{p} asociado a cada sentencia s, dentro de un documento d , para un usuario u, utilizando las palabras clave p del modelo a largo plazo, se obtiene calculando la similitud entre el vector de palabras clave y cada vector de cada una de las frases:

$$P_{sdu}^{p} = sim(s_{sd}, p_u) \tag{3.12}$$

donde:

- s_{sd} es el vector de pesos de términos de la sentencia s del documento d a resumir.

- p_u es el vector de pesos de términos correspondiente a las palabras clave del modelo a largo plazo del usuario u.
- sim es la medida de similitud del MEV.

A partir de la ecuación (3.12) se obtiene un ranking de sentencias para cada documento con respecto a la heurística de personalización utilizando las palabras clave del modelo a largo plazo.

3.5.2.3. Personalización utilizando el modelo a corto plazo

De manera similar se pueden obtener resúmenes personalizados utilizando el modelo a corto plazo o del usuario u. En este caso se calcula la similitud entre cada sentencia s y los términos de realimentación t_u, del modelo a corto plazo o, del usuario u. De esta forma, el valor P^o_{sdu} asociado a cada sentencia s, dentro de un documento d, para un usuario u, utilizando el modelo a corto plazo o, se obtiene calculando la similitud entre el vector de términos de realimentación y cada vector de cada una de las frases:

$$P^o_{sdu} = sim(s_{sd}, t_u) \tag{3.14}$$

donde:

- s_{sd} es el vector de pesos de términos de la sentencia s del documento d a resumir.
- t_u es el vector de pesos de términos correspondiente a los términos de realimentación del usuario u.
- sim es la medida de similitud del MEV.

A partir de la ecuación (3.14) se obtiene un ranking de sentencias para cada documento con respecto a la heurística de personalización utilizando el modelo a corto plazo.

3.5.2.4. Personalización utilizando la combinación de los modelos a corto y largo plazo

De manera similar a la selección, se pueden combinar los valores obtenidos para cada una de las partes del modelo de usuario para obtener un resumen personalizado que utilice el modelo de usuario completo. En este caso el valor P_{sdu} asignado a cada sentencia s, dentro del documento d, con respecto a un usuario u, se obtiene con la siguiente fórmula:

$$P_{sdu} = \frac{\eta P^p_{sdu} + \kappa P^o_{sdu}}{\eta + \lambda + \kappa} \tag{3.15}$$

donde las letras griegas η y κ muestran la importancia asignada a cada uno de los sistemas de referencia (η, para las palabras clave del modelo a largo plazo, κ, para modelo a largo plazo). Para que esta combinación sea significativa, la relevancia obtenida a partir de cada sistema de referencia debe ser normalizada con respecto a los mejores resultados para la colección de documentos que se esté utilizando.

El ratio entre η y κ representa la importancia relativa asignada dentro de la heurística de personalización al modelo a largo plazo frente al modelo a corto plazo. Los valores de η y κ actuarán como variables de control en los experimentos.

A partir de la ecuación (3.15) se obtiene un ranking de sentencias para cada documento con respecto a la heurística de personalización utilizando el modelo de usuario completo.

3.5.3. Combinación de las heurísticas genéricas y de personalización

Para obtener el valor Z_{sdu} asociado entre cada sentencia s, dentro de un documento d, para un usuario u, utilizando la combinación de las todas las heurísticas, se aplica la siguiente ecuación:

$$Z_{sdu} = \frac{\varphi A_{sd} + \gamma B_{sd} + \mu P_{sdu}}{\varphi + \gamma + \mu} \tag{3.16}$$

Los parámetros φ, γ y μ permiten el ajuste entre las diferentes heurísticas, dependiendo de si las heurísticas son genéricas (φ, γ) o de personalización (μ). De esta manera el valor de estos parámetros determina el grado de personalización de los resúmenes: si μ es 0, los resúmenes obtenidos serán genéricos, y para valores de μ mayores que 0, los resúmenes serán personalizados.

Para que esta combinación sea significativa, la relevancia obtenida a partir de cada sistema de referencia debe ser normalizada con respecto a los mejores resultados para la colección de documentos que se esté utilizando.

El valor entre φ+γ y μ, representa la importancia relativa asignada a las heurísticas de generación de resúmenes genéricos con respecto a la heurística de personalización. Los valores de φ, γ y μ actuarán como variables de control en los experimentos.

A partir de la ecuación (3.16) se obtiene un ranking de sentencias para cada documento con respecto a la combinación de todas las heurísticas propuestas.

3.6. Resumen de parámetros del sistema

El funcionamiento del sistema propuesto depende de varios parámetros de 2 tipos diferentes: unos con valores fijos y otros con valores ajustables. Para los experimentos presentados en este trabajo, algunos parámetros del sistema se han fijado empíricamente como valores válidos, para poder concentrar el estudio en el efecto de las alteraciones del resto de los parámetros. Los parámetros con valores fijos, representados como letras latinas mayúsculas, se muestran en la Tabla 3.2.

Proceso	Parámetro	Descripción	Rango	Valor
Adaptación	P	Peso de la realimentación positiva	0..1	0.9
	N	Peso de la realimentación negativa	0..1	0.9
	T	Peso del título		2
	V	Velocidad de cambio	0..1	0.8
	R	Número de términos de realimentación		10
	D	Decremento por día	0..1	0.1
Presentación	M	Num. documentos relevantes por día		10
	-	Heurística de posición		Tabla 3.1
	S	Número de palabras significativas		8
	L	Longitud del resumen en %		20

Tabla 3.2. Parámetros del sistema con valores fijos.

En el proceso de adaptación de contenidos se han fijado, para este trabajo, una serie de valores fijos para varios parámetros. En primer lugar, los parámetros P, N, T, V y R se les ha asignado los valores 0.9, 0.9, 2, 0.8 y 10, respectivamente, siguiendo el trabajo de [Nakashima&Nakamura97]. Por otro lado, se ha elegido un valor de D = 0.1. Este valor hace que los términos de realimentación duren como mucho 10 días, si no se vuelven a obtener de nuevos documentos realimentados.

En cuanto al proceso de presentación de resultados, en primer lugar se ha elegido un valor de M = 10. Se ha considerado este valor por ser un número razonable de documentos para no saturar al usuario, además fue el valor más utilizado en los experimentos preliminares realizados. Además se ha fijado S = 8 como número de palabras significativas por documento, ya que más allá de ese valor los pesos de las palabras empiezan a ser demasiado bajos para considerarlas como significativas. Por último, para este trabajo se ha elegido para L un valor de 20, ya que esta tasa de compresión suele ser lo suficientemente informativa como para sustituir al documento original [Morris *et al.* 92].

La Tabla 3.3 muestra el conjunto de parámetros sobre los que se van a estudiar distintas configuraciones a lo largo de los experimentos de los siguientes capítulos, éstos estarán representados por letras griegas en minúsculas. Todos estos parámetros toman valores entre 0 y 1.

Proceso	Parámetro	Descripción	Rango
Selección	α	Peso de las categorías propias	0 .. 1
	β	Peso de las categorías generales	0 .. 1
	χ	Peso de las palabras clave	0 .. 1
	ε	Peso del modelo a corto plazo	0 .. 1
Presentación	φ	Peso de la heurística de posición	0 .. 1
	γ	Peso de la heurística de palabras significativas	0 .. 1
	η	Peso de la heurística de personalización con palabras clave	0 .. 1
	λ	Peso de la heurística de personalización con categorías generales	0 .. 1
	κ	Peso de la heurística de personalización con corto plazo	0 .. 1
	μ	Peso de la heurística de personalización	0 .. 1

Tabla 3.3. Parámetros ajustables del sistema.

En todo caso, los distintos procesos se han parametrizado para permitir cualesquiera otros valores para los distintos parámetros, tanto para los valores fijos como para los ajustables.

3.7. Resumen y conclusiones del capítulo

Se ha descrito cómo se representan los documentos y el modelo de usuario y cómo se realizan cada uno de los procesos de personalización de contenidos: selección de contenidos, adaptación del modelo de usuario y presentación de resultados.

Los documentos se obtienen como la concatenación de título y cuerpo obtenidos de la página HTML bajada de la Web. La representación se realiza utilizando vectores de pesos de

términos, filtrados de palabras vacías, y en forma de raíces de palabras obtenidas mediante la utilización del algoritmo de Porter. El peso utilizado es tf · idf.

El modelo de usuario se subdivide en distintas partes para representar los intereses de los usuarios desde distintos puntos de vista. En primer lugar, se separan los intereses a largo y corto plazo. En segundo lugar, los intereses a largo plazo son representados en base a tres sistemas de referencia distintos: categorías propias, categorías generales y palabras clave. El modelo a corto plazo se representa como un conjunto de términos de realimentación.

El usuario puede editar y modificar su perfil a largo plazo, dando pesos a los distintos elementos de cada uno de los sistemas de referencia, así como a los sistemas de referencia en general. También puede añadir, modificar y borrar sus palabras clave.

La representación del modelo de usuario es diferente para cada sistema de referencia. Los pesos asignados a categorías propias y a categorías generales se representan en matrices con los valores introducidos por el usuario. Por otro lado, las categorías generales son representadas como vectores de pesos de términos obtenidos a partir de las descripciones breves de las páginas de las categorías de primer nivel de Yahoo! España. Por último, las palabras clave y los términos de realimentación se representan como vectores de pesos de términos, las primeras a partir de los pesos introducidos por el usuario, los segundos se obtienen en el proceso de adaptación del modelo de usuario.

Una contribución importante de esta tesis consiste en probar cuál es la mejor forma de combinar los distintos sistemas de referencia para obtener la mejor solución posible. Para lograr este propósito, cada uno de los 4 sistemas de referencia descritos (categorías propias, categorías generales, palabras clave y términos de realimentación) tiene un peso que indica su importancia en la obtención de los resultados finales. De esta forma, cada una de las 4 dimensiones puede ser definida y controlada durante los experimentos.

El proceso de selección se basa en la combinación de las similitudes obtenidas a partir de cada uno de los sistemas de referencia del modelo a largo plazo. Con respecto a las categorías propias, la selección es directa para cada documento, o pertenece no pertenece a una categoría propia seleccionada por el usuario. Con respecto a las palabras clave, la similitud se obtiene a partir de la fórmula del coseno entre los vectores que representan al documento y a las palabras clave. Por último, con respecto a las categorías generales hay que combinar las similitudes obtenidas, a partir de la fórmula del coseno, para cada una de las categorías generales seleccionadas por el usuario.

El proceso de adaptación del modelo de usuario genera/actualiza los términos de realimentación a partir de los términos obtenidos de los documentos sobre los que el usuario efectúa algún tipo de realimentación. Esta información se utiliza para hacer evolucionar el modelo del usuario dinámicamente y adaptarlo así a los cambios en los intereses del usuario que se puedan producir a lo largo del tiempo.

El proceso de presentación de resultados construye un documento Web con los documentos más relevantes seleccionados para cada usuario representados con título, autor, categoría propia, fuente, relevancia, resumen y enlace a la noticia completa. El resumen de cada documento se genera automáticamente y de manera personalizada, a partir de las frases más representativas obtenidas a partir de la combinación de una serie de heurísticas genéricas (posición y palabras temáticas) y de personalización. Esta última heurística se basa en la similitud de cada una de las sentencias del documento con el modelo de usuario (palabras clave, categorías generales y términos de realimentación).

La utilización de los resúmenes permite un ahorro de tiempo a los usuarios a la hora de detectar si un documento realmente le interesa sin tener que leerse el texto completo. Si

además el resumen está personalizado según sus intereses, el usuario tardará aún menos tiempo no sólo en decidir si le interesa o no, sino además en encontrar cuál es la información que realmente le interesa de esa noticia.

Los distintos procesos han sido parametrizados para permitir su adaptación a distintas configuraciones de personalización. En este trabajo, en particular, ha sido fijado el valor de algunos de estos parámetros para permitir el estudio del resto de ellos. En todo caso, se podrían variar también los valores de los parámetros fijos si se decidera experimentar con ellos.

Capítulo 4
METODOLOGÍA DE EVALUACIÓN

4.1. Introducción

En este tema se va a describir la metodología de evaluación utilizada para evaluar el modelo de sistema de personalización de contenidos propuesto en esta tesis. El esquema fundamental en el que se basa la personalización es el de un sistema de filtrado de información basado en un modelo de usuario. Esto lleva, por tanto, a pensar que lo más adecuado sea utilizar metodologías de evaluación utilizadas en el filtrado de información. Sin embargo, la personalización va a estar formada por los tres procesos que se describían en el tema anterior: selección de contenidos, adaptación del modelo de usuario y presentación de resultados. El funcionamiento distinto de cada uno de ellos hace que la metodología de evaluación tenga que ser adaptada para cada proceso.

En el apartado 4.2 se describirá cómo debe ser y cómo se puede obtener una colección de evaluación para poder evaluar los tres procesos de personalización.

En el apartado 4.3 se describirán los tipos de evaluación que se van a utilizar para juzgar los procesos de personalización.

En los apartados 4.4, 4.5 y 4.6 se describirán la metodología de evaluación utilizada en la selección, adaptación y presentación, respectivamente.

Por último, en el apartado 4.7 se mostrará un resumen y las conclusiones del capítulo.

4.2. Colección de evaluación

Es importante la existencia de una colección de evaluación para poder evaluar un sistema de manera sistemática, más allá de mostrar el funcionamiento para unos determinados casos concretos que no muestran la efectividad real. Además aporta un marco sobre el cual poder efectuar comparaciones entre distintas propuestas.

En general, una colección de evaluación está compuesta por un conjunto de documentos con una estructura más o menos similar, que suelen tratar de contenidos sobre un mismo dominio, (p.ej.: económico, periodístico, etc.), de un conjunto de tareas a realizar, y de los resultados asociados, que serán los juicios de relevancia determinados manualmente por expertos humanos. Por ejemplo, en recuperación de información, las tareas serían las consultas y los resultados serían los juicios de relevancia asociados a cada consulta.

Existen colecciones de evaluación para distintas tareas relacionadas con la clasificación de texto, algunas de ellas han surgido en el contexto de congresos tipo competición en los cuales se les suministra a los distintos participantes una misma colección y cada uno de ellos

trata de acercarse lo más posible a los resultados correctos. Quizá el contexto más conocido son las conferencias TREC[1], sobre recuperación de información. Otras conferencias de este tipo son: DUC[2] para generación de resúmenes, CLEF[3] para recuperación de información multilingüe, etc. Otro tipo de colecciones han surgido en otros contextos diferentes, normalmente como necesidad para la evaluación de algún nuevo tipo de sistema, pero suelen ser mucho menos completas.

Una de las fuentes más interesantes y de mayor accesibilidad para obtener documentos para construir una colección de evaluación es la Web [Grefenstette99]. En realidad es una fuente inagotable de documentos que se pueden utilizar como colección para evaluar un sistema. En todo caso, la idea de utilizar documentos de la Web como fuentes para una colección choca con la heterogeneidad extrema de los documentos HTML [Martínez *et al.* 91]: no comparten el mismo formato, ni aún procediendo de la misma fuente de información. Esto requiere un esfuerzo extra para preprocesar estos documentos y obtener la colección inicial "limpia", es decir, sin etiquetas y sólo con la información interesante.

Sin embargo, una colección no es sólo el conjunto de documentos, si no que también debe incluir las tareas a realizar y los juicios de relevancia. El conjunto de tareas debe elegirse de manera que represente un número de situaciones diferentes que permita juzgar el funcionamiento de un sistema de manera significativa. Por otro lado, la obtención de los juicios de relevancia supone un esfuerzo manual que debe ser realizado, en principio, por un experto humano, para obtener los documentos relevantes asociados a cada una de las tareas especificadas.

Las colecciones de evaluación para personalización tienen un inconveniente principal respecto a otras colecciones de evaluación: se necesitan juicios de relevancia distintos para cada usuario y para cada día. Esto es debido a que las tareas a realizar consisten en seleccionar los documentos más relevantes para cada usuario, cada día, y cada usuario tiene unas necesidades de información diferentes representadas en su modelo de usuario, además éstas pueden variar de un día a otro conforme a la información que reciban. Estos juicios pueden ser determinados por un experto humano basándose en el modelo de cada usuario o pueden ser elegidos por cada usuario, cada día, en función de sus necesidades de información en el momento de juzgar los documentos. Esta segunda opción es mucho más real, los usuarios determinan qué documentos son relevantes para sus intereses en el momento en el que los reciben, por tanto, utilizando sus necesidades de información actuales. En las colecciones habituales de clasificación de texto no ocurre esto porque los juicios son generales para todos los usuarios y son realizados por un experto humano que desconoce las necesidades de información concretas de distintos usuarios que pudieran querer realizar esas tareas en distintos momentos.

Además, como se pretende que esta personalización sea adaptativa con el paso del tiempo, es necesario recoger juicios de realimentación del usuario (positivo/negativo/nulo) sobre los documentos que recibe. De nuevo, este tipo de información no se recoge en las colecciones típicas de clasificación de texto porque la evaluación de las mismas se realiza en entornos estáticos que no evolucionan a lo largo del tiempo.

---

[1] Text REtrieval Conference (TREC). Home Page: http://trec.nist.gov/

[2] Document Understanding Conferences (DUC). Home Page: http://www-nlpir.nist.gov/projects/duc/intro.html

[3] Cross Language Evaluation Forum (CLEF) Home Page: http://clef.isti.cnr.it/

En principio, se debería haber utilizado una colección previamente definida por otros autores y que hubiese sido utilizada en experimentos parecidos. Esto permitiría la comparación de propuestas y la extracción de conclusiones asociadas a esa comparación. Desgraciadamente no existen colecciones generales asociadas a sistemas de personalización, precisamente por ser personalizadas: cada sistema utiliza sus propios modelos de usuario y sus juicios de relevancia asociados a dichos modelos. Por tanto, la construcción de la colección ha de hacerse procurando que sea lo más general posible para que se pueda utilizar por otros sistemas de personalización.

4.2.1. Proceso de construcción de una colección de evaluación

En este trabajo, la colección de evaluación está enfocada a sistemas de personalización de periódicos digitales, contiene información sobre modelos de usuario, sobre la relevancia de las noticias para los distintos usuarios y sobre la realimentación explícita producida por cada uno de ellos. Esta colección se puede aplicar a cualquier otro sistema de personalización que utilice técnicas similares de personalización o filtrado de información.

A partir de ahora, una vez fijado el dominio de aplicación a los periódicos digitales, hablaremos de secciones en lugar de categorías propias, y de categorías en lugar de categorías generales.

A continuación se describen brevemente los distintos pasos seguidos para la construcción de la colección:

En primer lugar, puesto que el dominio seleccionado es el de los periódicos electrónicos es necesario conectarse a la Web del periódico y bajarse los archivos correspondientes a las noticias de cada día. Se puede acceder a las noticias a partir de la página de cada sección, la cual almacena enlaces a todas las noticias contenidas en dicha sección [Martínez *et al.* 91].

Posteriormente, hay que extraer del formato HTML de los archivos bajados los elementos que sean interesantes para la colección. En general, la información interesante es la relacionada con el título, el autor, la sección, el texto del artículo y el enlace al texto completo. Esta información hay que almacenarla de manera permanente para ser procesada posteriormente.

El segundo paso de la construcción de la colección es obtener la definición de los intereses a largo plazo como modelos de usuario para poder empezar a ejecutar el sistema y obtener la primera personalización de contenidos. Los modelos de usuario iniciales son construidos por los usuarios el día o días anteriores al comienzo del experimento. Esto permite que el sistema disponga de información sobre los modelos para poder realizar la personalización del primer día. Estos perfiles iniciales contienen información sobre los intereses a largo plazo del usuario, esto es, secciones, categorías y palabras clave.

El tercer paso de la construcción de la colección es la obtención de los juicios de relevancia de cada uno de los usuarios respecto a todas las noticias, durante varios días. Para ello, se le envía cada día a cada usuario un mensaje de correo electrónico con un documento Web, generado por el proceso de presentación de resultados, que contiene como elementos de información todas las noticias del día. Cada usuario debe juzgar como relevante o no relevante cada una de esas noticias.

Con estos tres pasos se puede construir una colección de evaluación para evaluar la selección de contenidos.

Para poder juzgar la adaptación de contenidos necesitamos además los juicios sobre la realimentación del usuario sobre cada una de las noticias, durante todos los días que dure el experimento.

4.3. Tipos de evaluación

Se van a utilizar dos tipos de evaluación para juzgar los distintos procesos de personalización: una evaluación cuantitativa y una evaluación cualitativa.

La evaluación cuantitativa se va a basar en juicios de relevancia binarios asociados a los documentos. Esto es, en juicios sobre si los documentos son interesantes o no interesantes para cada usuario. Estos juicios pueden ser obtenidos, o bien por un experto a partir de los modelos de usuario y los documentos, o bien individualmente, por cada uno de los usuarios. El primer tipo de evaluación está más limitado puesto que sólo tiene en cuenta los intereses de los usuarios que están reflejados explícitamente en su modelo inicial

La evaluación cualitativa estará basada en las respuestas de los usuarios a uno o varios cuestionarios de evaluación donde se le preguntarán distintas cuestiones sobre su satisfacción en la utilización del sistema. De estas respuestas se pueden extraer conclusiones que pueden ser contrastadas con la evaluación cuantitativa.

Un aspecto que hay que tener en cuenta es que al medir el rendimiento diario, se miden tanto los efectos de la realimentación del modelo, como el efecto del cambio de noticias.

4.3.1. Evaluación cuantitativa

Para poder medir la efectividad de los distintos procesos de personalización es necesario fijar cuál va a ser la métrica que se va a utilizar. Los resultados que se van a obtener van a ser ranking de noticias para cada usuario. Estos resultados deben ser comparados con los juicios de relevancia binarios asociados. Esta comparación entre ranking de noticias y juicios de relevancia binarios (relevante/no relevante) sugiere la utilización como métricas de recall y precisión normalizados. La justificación de estas métricas es la comparación de rankings de noticias en lugar de grupos de noticias, es decir, no se observa simplemente si los X primeros documentos son relevantes o no, sino que se tiene en cuenta el orden en el cual aparecen en el ranking (no es lo mismo que estén los primeros entre los X o los últimos).

El recall y la precisión normalizados [Rocchio71] miden la efectividad del ranking en base al área (en una gráfica recall o precisión versus niveles de ranking) entre la mejor solución posible (documentos relevantes en las primeras posiciones) y la solución generada por el sistema a evaluar. Estas métricas se calculan con las ecuaciones (2.21) y (2.22). En los casos en los cuales coincide el mismo valor de relevancia para posiciones consecutivas del ranking, se toma como valor de la posición en el ranking de todos esos resultados iguales, el valor medio de las posiciones coincidentes [Salton&McGill83]. Este ajuste evita el problema de atribuir un orden aleatorio relativo a cada una de las posiciones que tienen el mismo valor de relevancia asociado.

Para obtener un único resultado final, hay que repetir los cálculos de recall y precisión normalizados, para todos los usuarios y para todos los días. La media de todos estos cálculos, constituirá la medida de la efectividad de la personalización realizada.

4.3.1.1. Significancia estadística

Para dos técnicas A y B, lo que se quiere demostrar es que con la técnica A se obtienen mejores resultados que con la técnica B (A > B) con respecto a un parámetro V. Para ello, se representará como V+ al número de veces que la técnica A da mejores resultados que la B, como V- al número de veces que la técnica B da mejores resultados que la A, y como empates al número de veces que ambas técnicas obtengan el mismo resultado. Esto será aplicado tanto al recall normalizado (V = R) como a la precisión normalizada (V = P).

Un resultado se considerará como estadísticamente significativo si pasa el *sign-test*, con parejas de valores, con un nivel de significación del 5% ($p <= 0.05$). La decisión de utilizar este tipo de test de significancia estadística está basada en el hecho de que no se hace ninguna suposición sobre la distribución subyacente, y que, dados los diferentes procesos de normalización que se aplican, a distintos niveles, se deben tener en cuenta los valores relativos en lugar de las magnitudes absolutas para establecer la significancia estadística [Salton&McGill83].

4.3.2. Evaluación cualitativa

La evaluación cualitativa está basada en las opiniones de los usuarios recogidas en las respuestas a uno o varios cuestionarios de evaluación suministrados a lo largo del funcionamiento del sistema. Este tipo de evaluación trata de obtener del usuario opiniones explícitas sobre distintas funcionalidades del sistema para que puedan ser contrastadas con los resultados obtenidos de la evaluación cuantitativa.

Esta información fue obtenida a partir de cuestionarios con preguntas específicas sobre los aspectos más relevantes del sistema. Los cuestionarios estaban formados por varias preguntas agrupadas por temas. En la mayoría de las preguntas había normalmente 5 opciones para elegir el grado de satisfacción. También había preguntas abiertas al final de los cuestionarios.

4.3.2.1. Cuestionarios de evaluación

La utilización de cuestionarios de evaluación está basada en diversos estudios empíricos, como el realizado en [Spink02] o en la investigación de [Johnson *et al.* 03], en la que se hace hincapié en los juicios de los usuarios a partir de varios indicadores de éxito en orden a relaciones significativas buscadas en la evaluación.

Los cuestionarios, están compuestos en su mayoría por preguntas cerradas, donde las respuestas se indican mediante la utilización de un rango de cinco niveles (muy alto, alto, regular, bajo, muy bajo) o mediante la dualidad Sí/No. Por otro lado, también se ofrece la posibilidad de respuestas abiertas para una serie de preguntas generales sobre el sistema.

En el Apéndice I se muestran los cuestionarios de evaluación inicial y de evaluación final utilizados en las distintas versiones de sistemas de personalización. Aunque los cuestionarios han ido evolucionando según se iban aplicando a los distintos sistemas de personalización, el esquema de construcción de los mismos ha sido muy similar.

Las preguntas giran alrededor de diversas facetas. Una primera referida a temas como el grado de satisfacción con los distintos elementos de la interfaz, la capacidad de personalización del sistema o la capacidad para acceder a la información y las posibilidades de búsqueda de información. Este apartado, que partió de otros trabajos [Pastor&Asensi99; Codina00], se ha ido modificando ligeramente en función de las diferentes versiones de los sistemas de

personalización y de los enfoques de las diferentes investigaciones realizadas. Se incluyen cuestiones sobre el grado de satisfacción con los componentes gráficos más importantes del sistema (por ejemplo, los diferentes iconos empleados), el grado en que el sistema es atractivo para el usuario, la facilidad para usar el sistema y su amigabilidad, así como preguntas sobre la gestión de los contenidos, por ejemplo en lo que se refiere a la calidad del sistema de enlaces, y por último, las formas de ayuda al usuario, por ejemplo en si existe una introducción o un tutorial del sistema.

En segundo lugar, se examinan las categorías y las secciones, utilizando criterios similares a los utilizados en [Slype91], como formas adecuadas de modelado del usuario. En lo que afecta a las secciones, en un primer momento se pregunta a los usuarios sobre la idoneidad de éstas a la hora de reflejar sus necesidades de información. En este sentido, también se pregunta al usuario si introduciría nuevas secciones para reflejar sus necesidades de información, en qué número, y cuáles. También se pregunta al usuario sobre el grado en que los documentos que son remitidos a partir de las secciones previamente seleccionadas aparecen antes que aquellos que no pertenecen. Finalmente, también se incluyen los posibles cambios en el uso de las secciones como forma de personalización, así como los momentos en que se produjeron.

A continuación el formulario se centra en la valoración de las categorías. Al igual que en el apartado anterior, interesa saber en qué grado las categorías que el sistema ofrece son las adecuadas para reflejar sus necesidades de información y si introduciría nuevas categorías. Si la respuesta es afirmativa, interesa conocer en qué medida y cuáles. Asimismo, se pregunta por el grado en que el sistema muestra noticias correspondientes a las categorías escogidas antes que las que no pertenecen a las categorías seleccionadas. Igualmente, se han incluido preguntas sobre la modificación potencial del uso del sistema de categorías.

Otro de los elementos de valoración es el de las palabras clave que utilizan los usuarios en el momento de fijar su modelo de usuario. En un primer momento, se pretende averiguar la valoración sobre esta opción para precisar las necesidades de información. Al mismo tiempo, también se pregunta acerca de la capacidad del sistema para mostrar las noticias correspondientes a las palabras clave escogidas antes que las noticias que no las contengan, y el nivel en que los documentos recuperados según las palabras introducidas se corresponden con sus necesidades de información. Igualmente, se pregunta por la transparencia que ofrece el sistema a la hora de recuperar documentos en función de las palabras introducidas, o si el usuario ha cambiado de palabras clave durante el uso del sistema.

En lo que afecta a los resúmenes, y partiendo del trabajo de Moreiro [Moreiro02], se pretende alcanzar la impresión global que tienen los usuarios acerca de su calidad, valorando su adecuación al contenido de las noticias, aportando los principales elementos informativos, o su adecuación a los perfiles de los usuarios, lo que se considera fundamental para acercarse a la consideración del satisfacción global del usuario respecto del sistema. Para aclarar esta cuestión completamente, también se pregunta sobre el grado en que los resúmenes estén bien construidos, y sean coherentes y claros. De igual forma, se realizan cuestiones a los usuarios acerca del nivel en que se evitan las redundancias o sobre los elementos informativos que no son representados en los resúmenes.

El siguiente área se ocupa de la selección de información y de la adaptación del sistema a lo largo del tiempo. Se analizan la validez de los rankings de noticias a partir del perfil seleccionado y el nivel en que el sistema se adapta con el tiempo a las necesidades de información de los usuarios, a sus juicios y a los cambios producidos en las necesidades de información durante la utilización del sistema.

El siguiente apartado sirve para determinar los criterios por los que el usuario entiende que una noticia es relevante o no. Así, antes de usar el sistema, se solicita al usuario que señale cuáles de los siguientes criterios son los que cree que va a aplicar a la hora de decidir si una noticia es relevante o no: la perspectiva y el planteamiento, la profundidad y la cantidad de información, el estilo, la relación con su perfil de usuario, la relación con su necesidad de información y con sus temas de interés, la capacidad que ofrece la noticia para añadir nuevo conocimiento frente a otros documentos relacionados, la novedad que la noticia supone, la utilidad de la misma, la cercanía y grado de motivación desde un punto de vista emocional, la cercanía desde un punto de vista geográfico, que, una vez leído el documento completo, sea lo que esperaba con anterioridad, la cercanía y familiaridad con el contenido expuesto y finalmente la cercanía y familiaridad con el lenguaje empleado. De igual forma, también se les pregunta por su nivel del interés por la información que va a recibir antes de usar el sistema.

Después de utilizar el sistema, a los usuarios se les pregunta, de los criterios anteriormente expuestos, por aquellos que en realidad han utilizado, por su nivel del interés por la información que ha recibido después de usar el sistema y por los aspectos relacionados con cada noticia que ha utilizado para decidir si una noticia es relevante o no: el título, la sección, la relevancia, el resumen o la noticia completa.

En definitiva, los criterios propuestos responden a diversos estudios [Barry&Schamber98; Su92, 98] preocupados por los criterios que en realidad los usuarios emplean para definir si un documento es relevante o no. Existen varios artículos [Greisdorf03; Spink *et al.* 98] que profundizan en el tema de la relevancia, apostando por ideas como la región de relevancia y por los niveles de relevancia. En realidad, los usuarios finales utilizan un amplio rango de criterios para tomar decisiones relacionadas con los elementos de información recuperados en un sistema de recuperación de información.

Lo siguiente que se solicita en el cuestionario es una valoración global del sistema en torno al nivel de satisfacción y a la confianza que tiene el usuario en el sistema después de trabajar con él. Este interés también se traduce en preguntas sobre el nivel en que ha resuelto sus necesidades de información, sobre el modo de trabajo que prefiere para definir su perfil, y sobre la medida en que el sistema personaliza bien la información.

También se incluye un conjunto de preguntas abiertas en las que el usuario-evaluador puede responder libremente sin atenerse a una valoración previamente fijada. Se pregunta acerca de las características que considera más importantes del sistema, de los elementos que echa en falta en el sistema, y acerca de su valoración sobre el grado de interactividad que permite. En este sentido, se pregunta acerca de las necesidades de información que tiene después de utilizar el sistema, acerca del tipo de información que le interesa más, no sólo desde un punto de vista temático, sino del tipo de documento (reportaje, crónica, editorial, etc.), también después de utilizar el sistema. También se pregunta si ha cambiado su perfil de usuario y las razones para hacerlo o no.

Finalmente, el último apartado es el que se dedica a los comentarios, donde el usuario puede exponer libremente cualquier aspecto que no haya sido cubierto en el resto del formulario.

4.4. Selección de contenidos

La selección de contenidos se refiere a la elección entre todos los documentos de entrada de aquellos que son más relevantes para un usuario dado, según su perfil o modelo.

4.4.1. Hipótesis

Las preguntas que se van a tratar de responder son las siguientes:

- ¿Cuál es el mejor mecanismo para seleccionar las noticias relevantes para cada usuario?
- Usando el modelo a largo plazo, ¿es mejor utilizar sólo las secciones, sólo las categorías, sólo las palabras clave o una combinación de las opciones anteriores?

Estas cuestiones dan lugar a las siguientes hipótesis:

H1. La selección de las noticias con respecto a los modelos de usuario, usando sólo el modelo a largo plazo, es mejor si se utiliza una combinación de todos los sistemas de referencia que si se utiliza cualquier otra combinación.

4.4.2. Experimentos

Para evaluar la selección de contenidos se va a someter al sistema a las distintas combinaciones de los parámetros asociados a este proceso, es decir, las distintas combinaciones de secciones, categorías y palabras clave, y de esta manera se obtendrán los resultados asociados a las distintas maneras de seleccionar las noticias más relevantes para cada usuario.

Hay que tener en cuenta que estos experimentos no sólo comparan las distintas combinaciones posibles de sistemas de referencia dentro del modelo a largo plazo presentado en este trabajo, sino que además se están comparando con técnicas que se utilizan en otros trabajos presentados en el estado del arte.

4.4.2.1. Experimento 1. Combinación de secciones, categorías y palabras clave, dentro del modelo a largo plazo.

Para comprobar la hipótesis H1, se van a aplicar a todos los usuarios los diferentes mecanismos de selección del modelo a largo plazo utilizado en este sistema (r_{du}^{l}). Estos diferentes mecanismos se configuran dando diferentes valores a los parámetros de la ecuación (3.8):

- S: sólo secciones ($\alpha \neq 0$, $\beta = 0$, $\chi = 0$).
- C: sólo categorías ($\alpha = 0$, $\beta \neq 0$, $\chi = 0$).
- P: sólo palabras clave ($\alpha = 0$, $\beta = 0$, $\chi \neq 0$).
- SC: combinación de secciones y categorías ($\alpha \neq 0$, $\beta \neq 0$, $\chi = 0$).
- SP: combinación de secciones y palabras clave ($\alpha \neq 0$, $\beta = 0$, $\chi \neq 0$).
- CP: combinación de categorías y palabras clave ($\alpha = 0$, $\beta \neq 0$, $\chi \neq 0$).
- SCP: combinación de secciones, categorías y palabras clave ($\alpha \neq 0$, $\beta \neq 0$, $\chi \neq 0$).

4.4.3. Evaluación

Para cada uno de las distintas configuraciones se obtendrán los valores correspondientes de recall y precisión normalizados, para cada usuario, para cada día. Para obtener un valor final, para cada configuración, se promediarán los resultados obteniendo una media final por usua-

rio y por día, que servirá como valor de efectividad del proceso de selección para cada una de las configuraciones establecidas para el modelo a largo plazo.

Una vez obtenidos los valores finales de recall y precisión normalizados, para cada configuración, éstos se compararán para determinar cuál es la mejor combinación de sistemas de referencia para la selección utilizando el modelo a largo plazo.

Se determinará cuantitativamente que una opción es mejor que otra si el valor asociado a la primera es mayor que el de la segunda. Adicionalmente se determinará si esa mejora es significativa mediante la aplicación del sign-test.

Por otro lado, la evaluación cualitativa basada en los cuestionarios de evaluación permitirá contrastar estos resultados cuantitativos en base a las opiniones emitidas por los usuarios respecto al proceso de selección.

4.5. Adaptación del modelo de usuario

La adaptación del modelo de usuario es necesaria por que las necesidades de los usuarios varían con el tiempo, principalmente como efecto de la interacción con la información que reciben [Belkin97; Billsus&Pazzani00]. Por esta razón el modelo de usuario debe ser capaz de adaptarse a esos cambios de interés. Esta adaptación se realiza mediante la interacción del usuario con el sistema, a través de la cual se obtiene información para la realimentación del perfil.

Lo que se va a evaluar es el efecto en la selección de contenidos debido a la adaptación del modelo de usuario. Por lo tanto, se volverá a utilizar recall-precisión normalizado.

4.5.1. Hipótesis

Las preguntas que se van a tratar de responder son las siguientes:

- ¿Es mejor utilizar un modelo estático a largo plazo, un modelo dinámico a corto plazo o una combinación de ambos?
- Suponiendo la combinación de ambos modelos como la mejor opción, ¿qué combinación de sistemas de referencia del modelo a largo plazo es mejor utilizar?

Estas cuestiones dan lugar a las siguientes hipótesis:

H2. La selección de las noticias con respecto a los modelos de usuario, es mejor si se utiliza una combinación de los modelos a corto y largo plazo, que si alguno de ellos se utiliza en solitario.

H3. La selección de las noticias con respecto a los modelos de usuario, utilizando una combinación de los modelos a corto y largo plazo, es mejor si se utiliza para el modelo a largo plazo la combinación de todos los sistemas de referencia que si se utiliza cualquier otra combinación.

4.5.2. Experimentos

Para evaluar la adaptación de contenidos se va a someter al sistema a las distintas posibilidades de este proceso para establecer que combinación de parámetros es la más adecuada. La información que se va a utilizar para este proceso es la realimentación que haya realizado

cada usuario sobre las noticias recibidas cada día. Se va a medir el efecto de la adaptación del modelo de usuario en la selección de contenidos, de tal forma que cuando mejor sea la adaptación de un determinado día, mejor será la selección del día siguiente.

De nuevo, estos experimentos no sólo comparan las distintas combinaciones posibles de sistemas de referencia para la adaptación de contenidos, sino que además se están comparando con técnicas que se utilizan en otros trabajos presentados en el estado del arte.

4.5.2.1. Experimento 2. Combinación de modelos a corto y largo plazo.

Para comprobar la segunda hipótesis (H2), se van a aplicar a todos los usuarios las diferentes combinaciones de los modelos a largo plazo y corto plazo, utilizando las distintas combinaciones de secciones, categorías y palabras clave para el modelo a largo plazo, para el proceso de selección (r_{du}). Estas diferentes combinaciones se configuran dando diferentes valores a los parámetros de la ecuación (3.10):

Se efectuarán las siguientes comparaciones:

- Modelo a largo plazo sólo con las secciones L(S) ($\alpha \neq 0, \beta = 0, \chi = 0, \varepsilon = 0$), modelo a corto plazo O ($\alpha = 0, \beta = 0, \chi = 0, \varepsilon \neq 0$) y combinación de ambos modelos L(S)O ($\alpha \neq 0, \beta = 0, \chi = 0, \varepsilon \neq 0$).
- Modelo a largo plazo sólo con las categorías L(C) ($\alpha = 0, \beta \neq 0, \chi = 0, \varepsilon = 0$), modelo a corto plazo O ($\alpha = 0, \beta = 0, \chi = 0, \varepsilon \neq 0$) y combinación de ambos modelos L(C)O ($\alpha = 0, \beta \neq 0, \chi = 0, \varepsilon \neq 0$).
- Modelo a largo plazo sólo con las palabras clave L(P) ($\alpha = 0, \beta = 0, \chi \neq 0, \varepsilon = 0$), modelo a corto plazo O ($\alpha = 0, \beta = 0, \chi = 0, \varepsilon \neq 0$) y combinación de ambos modelos L(P)O ($\alpha = 0, \beta = 0, \chi \neq 0, \varepsilon \neq 0$).
- Modelo a largo plazo con la combinación de secciones y categorías L(SC) ($\alpha \neq 0, \beta \neq 0, \chi = 0, \varepsilon = 0$), modelo a corto plazo O ($\alpha = 0, \beta = 0, \chi = 0, \varepsilon \neq 0$) y combinación de ambos modelos L(SC)O ($\alpha \neq 0, \beta \neq 0, \chi = 0, \varepsilon \neq 0$).
- Modelo a largo plazo sólo con la combinación de secciones y palabras clave L(SP) ($\alpha \neq 0, \beta = 0, \chi \neq 0, \varepsilon = 0$), modelo a corto plazo O ($\alpha = 0, \beta = 0, \chi = 0, \varepsilon \neq 0$) y combinación de ambos modelos L(SP)O ($\alpha \neq 0, \beta = 0, \chi \neq 0, \varepsilon \neq 0$).
- Modelo a largo plazo sólo con la combinación de categorías y palabras clave L(CP) ($\alpha = 0, \beta \neq 0, \chi \neq 0, \varepsilon = 0$), modelo a corto plazo O ($\alpha = 0, \beta = 0, \chi = 0, \varepsilon \neq 0$) y combinación de ambos modelos L(CP)O ($\alpha = 0, \beta \neq 0, \chi \neq 0, \varepsilon \neq 0$).
- Modelo a largo plazo sólo con la combinación de secciones, categorías y palabras clave L(SCP) ($\alpha \neq 0, \beta \neq 0, \chi \neq 0, \varepsilon = 0$), modelo a corto plazo O ($\alpha = 0, \beta = 0, \chi = 0, \varepsilon \neq 0$) y combinación de ambos modelos L(SCP)O ($\alpha \neq 0, \beta \neq 0, \chi \neq 0, \varepsilon \neq 0$).

En realidad el experimento 2 lo que determina es, para cada combinación de sistemas de referencia del largo plazo, si es mejor utilizarla sin el corto plazo, con el corto plazo, o es mejor utilizar sólo el corto plazo en ese caso.

4.5.2.2. Experimento 3. Combinación de largo y corto plazo, combinando secciones, categorías y palabras clave para el modelo a largo plazo.

Para comprobar la tercera hipótesis (H3), se van a aplicar a todos los usuarios las diferentes combinaciones, dentro del modelo a largo plazo, para combinar los modelos a largo y corto plazo, para el proceso de selección (r_{du}). Estas diferentes combinaciones se configuran dando diferentes valores a los parámetros de la ecuación (3.10), pero fijando $\varepsilon = 1$, por la utilización del corto plazo:

- L(S)O: sólo secciones para el largo plazo, y corto plazo ($\alpha \neq 0$, $\beta = 0$, $\chi = 0$, $\varepsilon = 1$).
- L(C)O: sólo categorías para el largo plazo, y corto plazo ($\alpha = 0$, $\beta \neq 0$, $\chi = 0$, $\varepsilon = 1$).
- L(P)O: sólo palabras clave para el largo plazo, y corto plazo ($\alpha = 0$, $\beta = 0$, $\chi \neq 0$, $\varepsilon = 1$).
- L(SC)O: secciones y generales para el largo plazo, y corto plazo ($\alpha \neq 0$, $\beta \neq 0$, $\chi = 0$, $\varepsilon = 1$).
- L(SP)O: secciones y palabras clave para el largo plazo, y corto plazo ($\alpha \neq 0$, $\beta = 0$, $\chi \neq 0$, $\varepsilon = 1$).
- L(CP)O: categorías y palabras clave para el largo plazo, y corto plazo ($\alpha = 0$, $\beta \neq 0$, $\chi \neq 0$, $\varepsilon = 1$).
- L(SCP)O: secciones, categorías y palabras clave para el largo plazo, y corto plazo ($\alpha \neq 0$, $\beta \neq 0$, $\chi \neq 0$, $\varepsilon = 1$).

En realidad el experimento 3 lo que determina es cual es la mejor opción entre las mejores opciones determinadas en el experimento 2, que van a ser las que correspondan a combinaciones de largo y corto plazo.

4.5.3. Evaluación

Puesto que lo que se evalúa es el efecto de la adaptación sobre la selección el proceso para obtener los valores finales de recall y precisión normalizados es el mismo que el descrito para la selección, salvo que no se utilizan los valores del primer día porque la adaptación se produce a partir del segundo día.

Una vez obtenidos los valores finales de recall y precisión normalizados, para cada configuración, se comparan para cada una de las comparaciones propuestas para determinar en cada caso cuál es la mejor elección. Adicionalmente se determinará, mediante la aplicación del sign-test, si esa elección es significativamente mejor que las otras.

Por otro lado, la evaluación cualitativa basada en los cuestionarios de evaluación permitirá contrastar estos resultados cuantitativos en base a las opiniones emitidas por los usuarios respecto al proceso de adaptación.

4.6. Presentación de resultados

La presentación de resultados consiste en, una vez seleccionados los documentos que le interesan a un usuario, generar un único contenido que contenga estos documentos como elementos de información. La generación de este contenido se personaliza teniendo en cuenta las distintas preferencias indicadas por el usuario en su perfil. Los contenidos generados se plasman en un documento web en formato HTML que puede ser enviado por correo electrónico o visualizado como una página web.

La principal aportación en la presentación de resultados es la generación de resúmenes personalizados, por lo tanto, la evaluación de este proceso se basará principalmente en la evaluación de estos resúmenes. Lo que se va a evaluar es cual es la pérdida de información significativa para un usuario cuando se le muestra un resumen en lugar del documento completo correspondiente. También se tendrán en cuenta las opiniones explícitas mostradas por los usuarios en los cuestionarios de evaluación.

Los resúmenes van a ser evaluados independientemente del resto de los experimentos utilizando una técnica de evaluación indirecta propuesta en [Maña *et al.* 99]. La técnica se basa en la suposición de que si un proceso de resumen de documentos es bueno, entonces el resumen obtenido debe retener tanto como fuera posible de la información que asegure una correcta selección con respecto al modelo de usuario. Este tipo de evaluación cuantitativa es característico de sistemas de extracción de resúmenes que se enmarcan dentro de un sistema más complejo que realiza otras funciones, como en sistemas de recuperación o de filtrado de información como los que se presentan en esta tesis.

Para cada usuario, se construye una versión personalizada de todas las noticias de la colección de evaluación resumiendo (usando el algoritmo que se pretende probar) cada uno de los documentos de la misma. Posteriormente, se aplica el mismo proceso de selección aplicado en el experimento 3 (en realidad, restringido al modelo de usuario correspondiente al usuario para el cual se ha generado la colección personalizada), pero usando la nueva versión personalizada (resumida) de la colección de evaluación. Los resúmenes generados se utilizan como datos de entrada en un proceso de selección equivalente al llevado a cabo para el conjunto completo de noticias del experimento 3. El mecanismo de selección empleado será el que obtenga mejores resultados en el experimento 3.

La hipótesis es que, si el proceso de generación de resúmenes empleado mantiene la información del documento que es relevante para un perfil de usuario, los resultados obtenidos deberían ser similares a los obtenidos para ese mismo usuario en el experimento 3, con las noticias completas, los cuales se tomarán como valores de referencia superiores. Cualquier desviación de este valor indica pérdida de información debido a "fallos" en la generación de los resúmenes, los cuales han producido una variación en el ranking obtenido para la colección de resúmenes con respecto al ranking obtenido con la colección de noticias completas.

Mediante la aplicación de un proceso similar para cada algoritmo de generación de resúmenes, este experimento debería mostrar una explícita, aunque indirecta, medida de la adecuación de la generación de resúmenes personalizados. Hay que tener en cuenta que la generación de resúmenes personalizados se entiende como un proceso de generación de resúmenes que preserva la información específica que es relevante para un modelo de usuario dado, más que la información que realmente resume el contenido de un documento.

Para esta evaluación, se han generado resúmenes para todas las noticias de un día, para todos los usuarios. Esto significa que se ha generado un resumen para cada noticia y cada

usuario, para todas las noticias de un día dado. Este proceso se ha repetido para todos los días del experimento.

4.6.1. Hipótesis

Las preguntas que van a tratar de responderse son las siguientes:

- Para la obtención de resúmenes personalizados utilizando únicamente la heurística de personalización, ¿es mejor utilizar un modelo estático a largo plazo, un modelo dinámico a corto plazo o una combinación de ambos?
- ¿Cuánto se pierde, en términos de información recibida por los usuarios, enviando un resumen de un documento en lugar del texto completo?
- ¿Qué tipo de resumen es mejor en ese sentido?

Estas cuestiones dan lugar a las siguientes hipótesis:

H4. Los resúmenes, obtenidos usando sólo la heurística de personalización, son mejores si se utiliza una combinación de los modelos a largo y corto plazo, que si alguno de ellos se utiliza en solitario.

H5. Los resúmenes obtenidos usando sólo la heurística de personalización son mejores, con respecto a la información seleccionada por los usuarios, que los resúmenes obtenidos extrayendo las primeras frases del texto del documento completo.

H6. Los resúmenes obtenidos utilizando únicamente la heurística de personalización son mejores con respecto a la información seleccionada por los usuarios, que los resúmenes obtenidos usando las heurísticas de generación de resúmenes genéricos y que los resúmenes obtenidos usando una combinación de heurísticas.

H7. Los resúmenes obtenidos usando sólo la heurística de personalización son peores que las noticias completas con respecto a la información seleccionada por los usuarios.

4.6.2. Experimentos

Para evaluar la presentación de resultados (generación de resúmenes) se va a someter al sistema a las distintas combinaciones de los parámetros asociados a este proceso, es decir, las distintas combinaciones de las heurísticas de generación de resúmenes, y de esta manera se obtendrán los resultados asociados a las distintas maneras de generar los resúmenes más adecuados para cada usuario.

La generación de resúmenes personalizados puede combinar los modelos a largo plazo (palabras clave) y a corto plazo (términos de realimentación). Aunque la combinación de estos modelos se ha mostrado mejor para la selección de contenidos, esta misma circunstancia no tiene porque repetirse con la generación de los resúmenes personalizados.

De nuevo, estos experimentos no sólo comparan las distintas combinaciones posibles de sistemas de referencia para la generación de resúmenes, sino que además se están comparando con técnicas que se utilizan en otros trabajos presentados en el estado del arte.

4.6.2.1. Experimento 4. Generación de resúmenes personalizados.

Para comprobar la cuarta hipótesis (H4) se van a evaluar para todos los usuarios todos los tipos de resúmenes diferentes que se pueden generar utilizando únicamente la heurística de personalización (P_{sdu}), es decir, las distintas combinaciones de la utilización de los modelos a largo y corto plazo para la generación de resúmenes personalizados utilizando sólo la heurística de personalización, teniendo en cuenta que se van a utilizar las palabras clave del modelo a largo plazo y los términos de realimentación del modelo a corto plazo. Las distintas posibilidades corresponden a la asignación de distintos valores a los parámetros de la ecuación (3.15):

- Rp(L): resumen personalizado utilizando sólo las palabras clave del modelo a largo plazo ($\eta \neq 0$, $\kappa = 0$).
- Rp(O): resumen personalizado utilizando sólo los términos de realimentación del modelo a corto plazo ($\eta = 0$, $\kappa \neq 0$).
- Rp(LO): resumen personalizado utilizando una combinación de las palabras clave del modelo a largo plazo y los términos de realimentación del modelo a corto plazo ($\eta \neq 0$, $\kappa \neq 0$).

Varias colecciones de evaluación diferentes se han generado para cada usuario, cada una de ellas consiste en el conjunto de resúmenes de las noticias originales obtenidos mediante la aplicación de cada uno de los métodos de generación de resúmenes personalizados indicados, es decir, habrá una colección para cada usuario de resúmenes personalizados usando el modelo a corto plazo, otra equivalente con el modelo a largo plazo y una tercera usando la combinación de ambos modelos.

Ya que la colección de resúmenes es diferente para cada usuario el proceso de selección debe hacerse usuario por usuario, es decir, en realidad se realizan tantos procesos de selección por cada tipo de resumen como usuarios haya.

4.6.2.2. Experimento 5. Combinación de heurísticas para la generación de resúmenes.

Para comprobar las hipótesis quinta, sexta y séptima (H5, H6 y H7) se van a evaluar para todos los usuarios todos los tipos de resúmenes diferentes que se pueden generar utilizando las distintas combinaciones de heurísticas (Z_{sdu}). Los resúmenes que se van a generar pueden ser de diferentes tipos, dependiendo de la heurística específica que se utilice para generarlos, esto corresponde a asignar distintos valores a los parámetros de las ecuaciones (3.15) y (3.16):

- Rb (resumen base): L% primeras frases de la noticia completa. Este resumen actuará como línea base de nuestros experimentos. En realidad sería igual que la heurística de posición si se extendieran los valores de manera decremental a todas las frases de un documento.
- Rg (resumen genérico): resumen obtenido usando las heurísticas de generación de resúmenes genéricos ($\varphi \neq 0$, $\gamma \neq 0$, $\mu = 0$).
- Rgp (resumen genérico-personalizado): resumen obtenido utilizando ambos tipos de heurísticas ($\varphi \neq 0$, $\gamma \neq 0$, $\mu \neq 0$, $\eta \neq 0$, $\kappa \neq 0$).

- Rp (resumen personalizado): resumen obtenido usando sólo la heurística de personalización (combinando los modelos a largo y corto plazo). ($\varphi = 0$, $\gamma = 0$, $\mu \neq 0$, $\eta \neq 0$, $\kappa \neq 0$). Es decir, corresponde a Rp(LO) del experimento 4.

Varias colecciones de evaluación diferentes se han generado para cada usuario, cada una de ellas consiste en el conjunto de resúmenes de las noticias originales obtenidos mediante la aplicación de cada uno de los métodos de generación de resúmenes indicados.

Las colecciones de resúmenes base y de resúmenes genéricos serán las mismas para todos los usuarios, puesto que no dependen de ningún aspecto del modelo de usuario, es decir, no están personalizadas. Esto permitirá evaluar a todos los usuarios a la vez.

Sin embargo, habrá una colección diferente por usuario de resúmenes personalizados y de resúmenes genérico-personalizados, puesto que éstas sí dependen del modelo del usuario. De esta manera, habrá que evaluar por separado a cada usuario con respecto a su colección de resúmenes personalizados, por un lado, y respecto a su colección de resúmenes genérico-personalizados, por otro.

4.6.3. Evaluación

La evaluación cuantitativa se va a basar en la selección de los resúmenes con respecto a un modelo de usuario, por lo tanto, el resultado final será un ranking de resúmenes de documentos ordenados según la relevancia del resumen con respecto al modelo de usuario. Este ranking se tiene que comparar con los juicios binarios del usuario para determinar la calidad de los resúmenes mediante esta evaluación indirecta. Por tanto, las métricas utilizadas van a ser las mismas que en los experimentos de selección y adaptación, es decir, recall y precisión normalizados.

Para cada uno de las distintas configuraciones se obtendrán los valores correspondientes de recall y precisión normalizados, para cada usuario, para cada día. Para obtener un valor final, para cada configuración, se promediarán los resultados obteniendo una media final por usuario y por día, que servirá como valor de efectividad del proceso de presentación de resultados para cada una de las configuraciones establecidas. Se utilizarán los mismos criterios de significancia estadística para determinar si los resultados son estadísticamente significativos.

Por otro lado, la evaluación cualitativa basada en los cuestionarios de evaluación permitirá contrastar estos resultados cuantitativos en base a las opiniones emitidas por los usuarios respecto al proceso de presentación de resultados.

4.7. Resumen y conclusiones del capítulo

En primer lugar, se han indicado las características que debe tener una colección de evaluación para un sistema de personalización de contenidos, para posteriormente indicar cuáles son los pasos que se deben seguir para construir una colección basada en un periódico electrónico. Después se han descrito las metodologías de evaluación utilizadas para cada uno de los procesos de personalización de contenidos aplicados en este trabajo.

La existencia de una colección de evaluación es fundamental para poder evaluar un sistema de manera sistemática, más allá de mostrar el funcionamiento para unos determinados casos concretos que no muestran la efectividad real. Además aporta un marco sobre el cual poder efectuar comparaciones entre distintas propuestas.

Una colección de evaluación está compuesta por un conjunto de documentos, de un conjunto de tareas a realizar, y de los juicios de relevancia determinados manualmente por expertos humanos. Las colecciones de evaluación para personalización tienen un inconveniente principal respecto a otras colecciones de evaluación: se necesitan juicios de relevancia distintos para cada usuario y para cada día. Esto es debido a que las tareas a realizar consisten en seleccionar los documentos más relevantes para cada usuario, cada día, y cada usuario tiene unas necesidades de información diferentes representadas en su modelo de usuario, además éstas pueden variar de un día a otro conforme a la información que reciban.

Se van a utilizar dos tipos de evaluación para juzgar los distintos procesos de personalización: una evaluación cuantitativa y una evaluación cualitativa. La evaluación cuantitativa se va a basar en juicios de relevancia binarios asociados a los documentos. La evaluación cualitativa estará basada en las respuestas de los usuarios a uno o varios cuestionarios de evaluación donde se le preguntarán distintas cuestiones sobre su satisfacción en la utilización del sistema.

Las métricas utilizadas en la evaluación cuantitativa son recall y precisión normalizados porque lo que se compara es como de efectivos son los rankings de noticias, es decir, no se observa simplemente si los X primeros documentos son relevantes, sino que se tiene en cuenta el orden en el cual aparecen en el ranking (no es lo mismo que estén los primeros entre los X o los últimos). Además para juzgar la significancia estadística se utiliza el *sign-test* con un nivel de significación del 5%.

Hay que tener en cuenta que al medir el rendimiento diario, se miden tanto los efectos de la realimentación del modelo, como el efecto del cambio de noticias.

La evaluación cualitativa está basada en las opiniones de los usuarios recogidas en las respuestas a uno o varios cuestionarios de evaluación suministrados a lo largo del funcionamiento del sistema. Este tipo de evaluación trata de obtener del usuario opiniones explícitas sobre distintas funcionalidades del sistema para que puedan ser contrastadas con los resultados obtenidos de la evaluación cuantitativa. Los cuestionarios contienen preguntas sobre distintos aspectos del sistema de personalización: grado de satisfacción con los distintos elementos de la interfaz, opiniones sobre la utilidad de las categorías, las secciones y las palabras clave, impresiones sobre los resúmenes, valoraciones sobre los procesos de selección y adaptación, criterios de relevancia utilizados, valoraciones globales y preguntas abiertas.

Para evaluar la selección de contenidos se somete al sistema a distintas configuraciones en los parámetros del proceso para determinar cuál es la mejor combinación de los mismos. Estos parámetros tienen que ver con la utilización de las distintas partes del modelo a largo plazo: secciones, categorías y palabras clave.

Para evaluar la adaptación de contenidos se añade al proceso de selección el modelo a corto plazo obtenido a partir de la realimentación del usuario y se estudian las distintas combinaciones de sistemas de referencia del modelo a largo plazo con el corto plazo.

La principal aportación en la presentación de resultados es la generación de resúmenes personalizados, por lo tanto, la evaluación de este proceso se basa en la evaluación de estos resúmenes. Lo que se evalúa es la pérdida de información significativa para un usuario cuando se le muestra un resumen en lugar del documento completo correspondiente.

Para cada usuario, se construye una versión personalizada de todas las noticias de la colección de evaluación resumiendo (usando el algoritmo que se pretende probar) cada uno de los documentos de la misma. Posteriormente, se aplica el mismo proceso de selección aplicado en el experimento 3 (en realidad, restringido al modelo de usuario correspondiente al usuario para el cual se ha generado la colección personalizada), pero usando la nueva versión

personalizada (resumida) de la colección de evaluación. Los resúmenes generados se utilizan como datos de entrada en un proceso de selección equivalente al llevado a cabo para el conjunto completo de noticias del experimento 3. El mecanismo de selección empleado es el que obtiene mejores resultados en el experimento 3.

La hipótesis es que, si el proceso de generación de resúmenes empleado mantiene la información del documento que es relevante para un perfil de usuario, los resultados obtenidos deberían ser similares a los obtenidos para ese mismo usuario en el experimento 3, con las noticias completas, los cuales se tomarán como valores de referencia superiores. Cualquier desviación de este valor indica pérdida de información debido a "fallos" en la generación de los resúmenes, los cuales han producido una variación en el ranking obtenido para la colección de resúmenes con respecto al ranking obtenido con la colección de noticias completas.

Por tanto, para evaluar la presentación de resultados se somete al sistema a distintas configuraciones en los parámetros del proceso para determinar cuál es la mejor combinación de los mismos. Estos parámetros tienen que ver con la utilización de distintas heurísticas para generar los resúmenes: posición, palabras temáticas y personalización.

También hay que tener en cuenta las opiniones explícitas mostradas por los usuarios en los cuestionarios de evaluación, sobre los distintos procesos de personalización.

Capítulo 5
COLECCIONES DE EVALUACIÓN

5.1. Introducción

En este capítulo se van a explicar las distintas colecciones de evaluación generadas, en orden cronológico, que se han utilizado para evaluar los distintos sistemas de personalización utilizados en esta tesis doctoral. La evaluación de dichos sistemas se presentará en el siguiente capítulo.

En el apartado 5.2 se describirán varias minicolecciones preliminares que se construyeron para realizar unos primeros experimentos con los procesos de selección de contenidos y presentación de resultados.

En el apartado 5.3 se describirán los detalles técnicos de la construcción de las colecciones de evaluación generadas. El proceso seguido para construirlas es el que se describió en el apartado 4.2.1: obtención de noticias, de modelos de usuario, de juicios de relevancia y de juicios de realimentación.

En el apartado 5.4 se describirá la primera colección de evaluación generada para evaluar los tres procesos de personalización: selección, adaptación y presentación. El sistema utilizado para generar esta colección es el sistema de personalización 1.0. Esta colección [Díaz *et al.* 03] incluye evaluaciones de 11 usuarios reales sobre todas las noticias de cada día, durante 5 días.

En el apartado 5.5 se describirá la segunda colección [Díaz&Gervás04c], que también se generó para evaluar los tres procesos de personalización. El sistema utilizado fue el sistema de personalización 2.0. En este caso, el número de días utilizado fue 14, y el número de usuarios reales fue 106.

Por último, en el apartado 5.6 se mostrará un resumen y las conclusiones del capítulo.

5.2. Minicolecciones preliminares

Se generaron tres minicolecciones de evaluación para realizar unos experimentos preliminares sobre selección de contenidos y presentación de resultados. Estos experimentos se describen en el Capítulo 6.

5.2.1. Primera minicolección

La primera minicolección de evaluación se generó durante los días laborables del 28 de Enero al 10 de Febrero de 2000. El número de noticias bajadas cada día fue aproximadamente 100. Se utilizaron las siguientes 8 secciones del periódico ABC: nacional, internacional, de-

portes, economía, sociedad, cultura, gente y opinión. También hay que resaltar que por fallos del sistema los usuarios no recibieron noticias los días 29 y 30 de Enero, ni el 3 de Febrero.

Esta minicolección se utilizó para realizar una primera evaluación del proceso de selección de contenidos. El mensaje que recibía el usuario contenía solamente el número de noticias que él había seleccionado en su modelo.

El mecanismo utilizado para dar difusión al sistema fue el envío de correos electrónicos solicitando participación, a todos los profesores de la Ingeniería en Informática y de Periodismo, de la Universidad Europea de Madrid. Además se permitió cualquier tipo de difusión adicional.

En el mensaje de difusión se indicó a los posibles usuarios en que consistía el proceso de evaluación. Se eligieron 3 días durante los cuales cada usuario debía realizar una evaluación exhaustiva del sistema, esto es, debía indicar cuántas noticias de las que recibía eran relevantes y cuántas noticias de las que no recibía eran relevantes. Estos días fueron el 1, 7 y 9 de Febrero. Hay que resaltar que el usuario tenía que revisar en la web del periódico todas las noticias del día para determinar cuántas de las que no recibía eran relevantes. Finalmente, los usuarios tenían que rellenar un cuestionario de evaluación donde debían indicar los datos de la evaluación exhaustiva, así como distintas opiniones sobre el sistema recogidas en un cuestionario de evaluación (Apéndice I) que se les enviaba en el propio mensaje de difusión.

Se dispuso de 44 modelos de usuario correspondientes a 44 personas diferentes. En esta colección se separó a los usuarios en 4 grupos: colaboradores (A), investigadores (B), profesores de informática y periodismo (C) y usuarios externos (D). Los perfiles de usuario contienen información sobre los intereses a largo plazo del usuario, esto es, secciones, categorías y palabras clave.

Algunos usuarios seleccionaron un método (secciones por ejemplo) pero no seleccionaron ninguna posibilidad para él (no marcaron ninguna sección específica). Esto produjo modelos de usuario vacíos. Esto le ocurrió a 14 usuarios (31.8%), todos los cuales sólo eligieron secciones. Este es un problema a resolver para las siguientes versiones del sistema. Estos usuarios han sido eliminados de la evaluación del sistema para evitar resultados incorrectos en la evaluación, por lo tanto, el número de usuarios en la colección de evaluación fue 30.

Hay que resaltar que para la obtención de los juicios de relevancia de cada uno de los usuarios se eligieron 3 días durante los cuales cada usuario debía realizar una evaluación exhaustiva del sistema. Estos días fueron el 1, 7 y 9 de Febrero. Sin embargo, sólo se indicó a los usuarios que mostraran el número de noticias relevantes y no cuáles eran relevantes. Por lo tanto, esta información no se pudo utilizar para construir la colección de evaluación.

Los juicios de relevancia se obtuvieron examinando cada uno de los modelos de usuario con respecto a todas las noticias de un día y determinando cuáles eran las relevantes. Este examen no lo hacen cada uno de los usuarios sino que lo realiza un experto humano para todos los modelos. Este proceso sólo se realizó para el último de los días que los usuarios tenían que realizar la evaluación exhaustiva, es decir, el 9 de Febrero. Ese día hubo 109 noticias.

Los datos concretos sobre la primera minicolección de evaluación aparecen en [Díaz *et al.* 00a].

5.2.2. Segunda minicolección

La segunda minicolección se generó a partir de la minicolección anterior añadiendo 36 nuevos perfiles a los 30 perfiles no vacíos anteriores. Estos nuevos pseudoperfiles no corresponden a usuarios reales y fueron generados para representar distintas configuraciones de los modelos de usuario y obtener así unos resultados más generales sobre el sistema. Se construyó un perfil distinto para cada sección (8) y para cada categoría (14), para examinar más detenidamente el funcionamiento de cada uno de los sistemas de clasificación. Los restantes 14 perfiles contenían combinaciones de los tres sistemas de clasificación, secciones y categorías, secciones y palabras clave, categorías y palabras clave y sólo palabras clave. Además en estos perfiles se suministraron diferentes pesos a secciones, categorías y palabras clave, tanto a cada uno en particular como a los valores generales asociados a cada sistema de clasificación.

Los juicios de relevancia utilizados se refieren al mismo día que fue utilizado en la primera minicolección. Para ello, se examina cada uno de los modelos de usuario con respecto a todas las noticias de ese día (109) y se determina cuáles son las relevantes. Además se fija 20 como límite superior en las noticias recibidas para todos los usuarios, esto se hace así para evaluar la precisión y el recall en los mismos niveles para todos los usuarios. Este examen no lo hacen cada uno de los usuarios sino que lo realiza un único experto humano para todos los modelos. Hay que resaltar que no sólo se mira cuántas noticias hay relevantes o no relevantes sino que se marcan cada una de ellas para realizar distintas evaluaciones a distintos niveles de recall.

Los datos concretos sobre la segunda minicolección de evaluación aparecen en [Díaz *et al.* 01a].

5.2.3. Tercera minicolección

La tercera minicolección se generó como un primer intento para la evaluación de la presentación de resultados. Se trataba de generar resúmenes personalizados para varios usuarios utilizando las heurísticas descritas en el capítulo 3, teniendo en cuenta que en este experimento no existía la posibilidad de adaptación del modelo de usuario.

Se seleccionaron 3 perfiles no triviales de la primera minicolección de evaluación, es decir, correspondientes a usuarios reales y se utilizaron los juicios de relevancia de cada uno de los usuarios ya obtenidos en esa minicolección.

Los detalles de la tercera minicolección de evaluación aparecen en [Acero *et al.* 01].

5.3. Detalles técnicos de la construcción de las colecciones de evaluación generadas

A continuación se indican los detalles concretos utilizados en el proceso de construcción de las colecciones de evaluación utilizadas en esta tesis. El proceso seguido para construirlas es el que se describió en el apartado 4.2.1: obtención de noticias, de modelos de usuario, de juicios de relevancia y de juicios de realimentación. Algunos de los detalles indicados también se utilizaron en la construcción de las minicolecciones preliminares

En primer lugar, se ha construido un programa que se conecta diariamente a la Web de ABC y se baja todas las noticias del día. En particular se conecta con las páginas de las sec-

ciones utilizadas y se baja todas las noticias que aparecen en las mismas. En nuestro caso, se utilizan las siguientes 8 secciones del periódico: nacional, internacional, deportes, economía, sociedad, cultura, gente y opinión. El número medio de noticias por día es aproximadamente 100.

Se extraen de los ficheros HTML el título, el autor, la sección, el texto del artículo y el enlace al texto completo. Hay que resaltar que el tratamiento de esta información no es trivial, puesto que aunque las páginas web se generan automáticamente en el periódico mediante motores ASP, el aspecto final del código HTML generado es bastante caótico. Afortunadamente, la búsqueda de las etiquetas título, autor y cuerpo permite la localización de la información relevante sin muchas dificultades. La etiqueta entradilla (resumen generado por el editor) fue utilizada en experimentos preliminares, pero dejó de utilizarse porque no aparecía en todas las noticias. Este esquema no es válido para otros periódicos electrónicos ya que la forma de generar la información es diferente.

El formato final en el que se almacenan las noticias es como ficheros de texto: una línea para el título, otra para los autores, otra para la entradilla y el resto para el cuerpo de la noticia. En los casos en los que no se ha utilizado entradilla, la tercera línea está en blanco. La utilización de lenguajes de marcado (SGML, HTML, XML, ...) para la definición y almacenamiento de estas colecciones quizás sería mejor solución para estandarizar la colección [Martínez *et al.* 91], pero los archivos de texto son igualmente independientes de la plataforma y el contenido es suficientemente sencillo como para no necesitar un etiquetado del mismo.

El formato final en el que se almacenan los modelos de usuario es como *scripts* de SQL que almacenan en distintas tablas los distintos intereses de los modelos de usuario. Previamente a la ejecución de estos *scripts* hay que ejecutar otros que crean la base de datos y la inicializan con la información necesaria: secciones, categorías y representación de las categorías. Los detalles sobre la base de datos que maneja el sistema se describen en el Apéndice V.

El formato final en el que se almacenan los juicios de relevancia es como ficheros de texto que contienen tantas líneas como noticias y en cada línea, juicio de relevancia, espacio y noticia. El juicio de relevancia puede ser 0, no relevante ó 1, relevante. La noticia se representa como sección/fichero. Existe un fichero de juicios de relevancia por usuario y por día. Esta información se almacena en un árbol de directorios que en el primer nivel contiene tantos directorios como días y dentro de cada subdirectorio, tantos directorios como usuarios haya en el sistema.

El usuario dispone de dos iconos de realimentación (realimentación positiva/realimentación negativa) asociados a cada noticia a través de los cuales puede introducir sus juicios de realimentación. En realidad, el usuario puede elegir entre tres opciones a la hora de introducir estos juicios, asociados a las acciones de realimentar positivamente, no realimentar o realimentar negativamente. Además se indica a los usuarios que sus juicios de realimentación no se deben basar sólo en su perfil a largo plazo sino que deben reflejar sus necesidades de información, precisamente los intereses que no sean acordes a su modelo a largo plazo deberían ser capturados por el modelo a corto plazo.

En realidad, los juicios de relevancia de las noticias se han obtenido a partir de la realimentación del usuario puesto que la información introducida en ese proceso sirve para determinar la relevancia de las noticias para cada usuario. Sin embargo, puesto que lo que interesa es tener juicios de relevancia binarios se presentan dos posibilidades: se pueden tomar como noticias relevantes sólo las que reciban realimentación positiva, o considerar también como relevantes las que son indiferentes para el usuario. En este caso, se ha optado por el primer caso: sólo son relevantes las noticias con realimentación positiva.

La equivalencia establecida puede ser discutible porque un usuario puede interpretar la realimentación como una intención de querer saber más de una noticia y por tanto, no realimentar una noticia que es relevante para sus intereses pero de la cual no desea conocer más información. Esta interpretación haría que el mecanismo utilizado para la obtención de los juicios de relevancia no fuera válido. En ese caso, además, serían necesarios dos procesos distintos de evaluación de las noticias, uno para la obtención de los juicios de relevancia y otro distinto para la obtención de los juicios de realimentación. Esta duplicación de procesos saturaría aún más a los usuarios de lo que lo hace el hecho de tener que emitir juicios una sola vez sobre todas las noticias de cada día. Además cuando se realiza la evaluación se indica a los usuarios que el significado de la pulsación del icono de realimentación positiva es que la noticia le interesa y que, por tanto, desea recibir más información relacionada con ella. Por otro lado, la realimentación negativa implica, no sólo que no le interesa, sino que además no desea recibir ninguna información relacionada con ella. Por tanto, el esquena de obtención de juicios de relevancia y realimentación simultáneos es válido para la construcción de la colección de evaluación.

El formato final en el que se almacenan los juicios de realimentación es como ficheros de texto similares a los de los juicios de relevancia, salvo que en lugar de juicios de relevancia se almacenan juicios de realimentación asociados a noticias. En este caso, cada juicio de realimentación puede tener valor 1, 0 ó –1, según represente realimentación positiva, no realimentación o realimentación negativa. El árbol de directorios es el mismo, por lo tanto, se almacena en el mismo subdirectorio los juicios de relevancia y realimentación asociados a un determinado usuario durante un determinado día.

Las distintas colecciones generadas se presentan en los siguientes apartados.

5.4. Colección de evaluación 1.0

Se obtuvieron las noticias del período comprendido entre los días del 6 al 10 de Mayo de 2002. El número de noticias bajadas cada día fue 128, 104, 87, 98 y 102. Se utilizaron las siguientes 8 secciones del periódico: nacional, internacional, deportes, economía, sociedad, cultura, gente y opinión. El número total de noticias fue 519. Un ejemplo de noticia aparece en el Apéndice II.

Esta colección [Díaz *et al.* 03] se utilizó en el sistema de personalización 1.0, en el cual se implementan los tres procesos de personalización, pero el modelo de usuario a largo plazo se simplifica mediante la eliminación de las categorías.

El segundo paso de la construcción de la colección es obtener la definición de los intereses a largo plazo de varios modelos de usuario para poder empezar a ejecutar el sistema y obtener la primera personalización de contenidos.

El mecanismo utilizado para dar difusión al sistema fue el envío de correos electrónicos solicitando participación, a todos los profesores de la Ingeniería Técnica de Informática de Sistemas, y a algún profesor de otras carreras, del CES Felipe II de Aranjuez. Adicionalmente se permitió cualquier tipo de difusión adicional. Se obtuvieron 11 modelos de usuario correspondientes a 11 personas diferentes. Estos perfiles corresponden a 9 profesores de informática, una profesora de empresariales y una persona externa a la universidad.

En el mensaje de difusión se indicó a los posibles usuarios en que consistía el proceso de evaluación. En este caso, los usuarios recibían un mensaje cada día con la siguiente información asociada a cada noticia: título, autor, sección, resumen del artículo, enlace a la noticia

completa e iconos de realimentación. En este caso, el resumen está formado por las primeras frases del artículo. Los usuarios debían emitir sus evaluaciones interaccionando con las noticias mediante la pulsación o no de los iconos de realimentación.

Los perfiles iniciales contienen información sobre los intereses a largo plazo del usuario, que en este caso son sólo secciones y palabras clave. Estos intereses, tanto para secciones, como para palabras clave, son introducidos por el usuario manualmente utilizando una escala de 0 a 3. Un ejemplo de modelo de usuario aparece en el Apéndice II.

Usuario	Secciones	PalabrasClave
0	2	3
1	4	3
2	4	0
3	7	5
4	4	4
5	8	3
6	8	6
7	7	7
8	2	4
9	5	5
10	6	3
media	5.2	3.9

Tabla 5.1. Número de secciones y palabras clave elegidas por los usuarios en la primera colección de evaluación.

La Tabla 5.1 muestra el número de secciones y palabras clave elegidos por los usuarios en sus perfiles iniciales, esto es, en sus modelos a largo plazo. En general, los usuarios eligieron varias secciones y varias palabras clave. Estos modelos contenían, en media, 5.2 secciones y 3.9 palabras clave. Es interesante resaltar que un usuario (número 2) no seleccionó ninguna palabra clave, mientras que el resto introdujeron 3 como mínimo y 7 como máximo. También es un dato significativo que 2 usuarios (números 5 y 6) eligieran todas las secciones como relevantes, mientras que el mínimo de secciones elegidas fue 2.

La Tabla 5.2 muestra los pesos asignados a cada una de las secciones. Se puede observar que las secciones más importantes para los usuarios son internacional y cultura, un 82% de los usuarios (9 de los 11) las eligieron, aunque la sección de cultura tiene un mayor promedio (0.69) en los valores de los pesos asignados por los usuarios. Las menos elegidas son opinión y gente, 45% de los usuarios (5 de los 11). Sin embargo, la de mayor peso promedio es economía (0.89) y la de menor peso promedio es deportes (0.57). También se puede resaltar que los usuarios variaron sus pesos entre las distintas secciones, es decir, no utilizaron sólo el criterio me interesa/no me interesa, lo cual demuestra que el sistema de asignación de pesos es atractivo para los usuarios.

Usuario	Opinión	Nacional	Internacional	Economía	Sociedad	Cultura	Deportes	Gente
0	0	1	0	0	0	0	1	0
1	0	0.33	0.66	1	0	0	0	0.33
2	0	0	1	0	0	1	0	1
3	0.33	0.66	0.66	1	0.33	1	1	0
4	1	0	0	0	1	1	0.66	0
5	0.66	0.66	0.66	1	0.66	0.33	0.33	0.33
6	0.66	0.66	0.66	0.66	0.33	1	0.33	0.66
7	0.33	0.66	1	0.66	0.66	1	0.33	0
8	0	0	0.66	0	0	0.66	0	0
9	0	0.66	0.66	0	1	0.66	0	1
10	0	0.33	0.66	1	1	0.66	0.33	0
Valores seleccionados	5	8	9	6	7	9	7	5
Media	0.27	0.45	0.60	0.48	0.45	0.66	0.36	0.30
Media seleccionados	0.60	0.62	0.74	0.89	0.71	0.81	0.57	0.66

Tabla 5.2. Pesos asignados a las distintas secciones por los usuarios.

En cuanto a los pesos de las palabras clave cabe destacar que todos los usuarios asignaron peso 1 a sus palabras clave excepto 3 usuarios que escogieron para su último término peso 0.66. Es curioso resaltar que de las 43 palabras clave introducidas por todos los usuarios 12 fueron nombres propios y casi todas los demás se referían a temas generales, p. ej.: literatura, poesía, teatro, cine, informática, seguros, finanzas, educación, universidad, investigación, ciencia, historia, biología, monarquías, etc. Otro dato es que sólo se repiten dos palabras entre todos los usuarios: literatura y poesía, 2 veces.

El tercer paso de la construcción de la colección es la obtención de los juicios de relevancia de cada uno de los usuarios respecto a todas las noticias, durante los 5 días. Para ello, se le envía cada día a cada usuario un mensaje de correo electrónico con la siguiente información asociada a cada noticia: título, autor, sección, resumen del artículo, enlace a la noticia completa e iconos de realimentación.

Cada usuario puede pulsar en uno de los iconos de realimentación o no pulsar en ninguno, para emitir su juicio, esta pulsación invoca a un programa JSP que actualiza la información correspondiente a los juicios de cada usuario. Los juicios de relevancia de las noticias se obtienen a partir de la realimentación del usuario. Hay que tener en cuenta que los juicios de relevancia finales que se manejan son juicios binarios, relevante/no relevante, por ello se toman como relevantes sólo las noticias con realimentación positiva, mientras que se toman como no relevantes aquéllas con realimentación negativa o sobre las que no se haya efectuado ninguna realimentación.

Además se indica a los usuarios que sus juicios de realimentación no se deben basar sólo en su perfil a largo plazo sino que deben reflejar sus necesidades de información, precisamente los intereses que no sean acordes a su modelo a largo plazo deberían ser capturados por el modelo a corto plazo.

La Tabla 5.3 muestra el número de noticias indicadas como relevantes por cada uno de los usuarios durante cada uno de los 5 días que duró el experimento. El número medio de noticias relevantes por día varía entre 25.6 y 33.6, con 28 como valor medio. Los valores sufren una variación considerable, entre 5 noticias relevantes (10-Mayo, usuario 1) y 80 noticias relevantes (7-Mayo, usuario 8). El número total de evaluaciones realizadas fue de 5709 (11 usuarios, 519 noticias por usuario durante los 5 días).

Hay una diferencia significativa entre un conjunto de usuarios más exigentes, que juzgan pocas noticias como relevantes, y los usuarios que juzgan muchas noticias como relevantes. El primer grupo lo constituyen los usuarios 0, 1, 2, 3, 4, 7 y 10, que juzgan, en media, 16.6, 11, 11.2, 15.2, 14.6, 22.8 y 17.6 noticias como relevantes. El segundo grupo es el formado por los usuarios 5, 6, 8 y 9, que juzgan, en media, 32.4, 56.2, 68 y 42.8 noticias como relevantes.

Día	6-5	7-5	8-5	9-5	10-5	Media
NumNoticias	128	104	87	98	102	103.8
Usuario						
0	18	16	11	17	21	16.6
1	23	8	10	9	5	11.0
2	16	8	10	8	14	11.2
3	20	15	9	10	22	15.2
4	30	12	16	6	9	14.6
5	30	35	33	27	37	32.4
6	72	72	43	53	41	56.2
7	39	21	20	24	10	22.8
8	77	80	72	60	51	68.0
9	23	47	41	42	61	42.8
10	22	16	17	17	16	17.6
Media	33.6	30.0	25.6	24.8	26.1	28.0

Tabla 5.3. Número de noticias relevantes por usuario y por día.

El cuarto paso de la construcción de la colección es la obtención de los juicios sobre la realimentación del usuario sobre cada una de las noticias, durante los 5 días. Esta información se obtiene igualmente de las mismas pulsaciones de los usuarios sobre los iconos de realimentación de las que se obtienen los juicios de relevancia, sólo que ahora se interpretan las tres posibilidades: realimentación positiva, no realimentación o realimentación negativa.

En realidad, en la construcción de esta colección se cometió un pequeño error porque se le indicó a los usuarios que sólo realimentaran negativamente sobre las primeras 20 noticias que recibieran, aunque podían realimentar positivamente sobre todas las noticias. La justificación de esta decisión fue que el resto de noticias difícilmente iban a ser seleccionadas por su perfil de usuario entre las M primeras y por tanto, el impacto en el proceso de adaptación del modelo de usuario iba a ser mínimo o nulo. Esta decisión, en todo caso, no afecta a la obtención de los juicios de relevancia porque si el usuario no realimenta positivamente es que la noticia no le interesa. En cualquier caso la decisión se tomó para no saturar excesivamente a los usuarios a la hora de introducir sus juicios de relevancia.

La Tabla 5.4 muestra el número de noticias realimentadas, tanto positiva como negativamente (R+/R-), por cada uno de los usuarios durante cada uno de los 5 días que duró el experimento. El usuario 2 es el más exigente, sólo realimenta positivamente 11 noticias de entre todas, en media, y sin embargo, realimenta negativamente una media de 17.4 noticias de entre las 20 primeras. El 8 es el menos exigente, no realimenta negativamente ninguna noticia y realimenta positivamente 68, en media. El resto de los usuarios realimentan negativamente entre 2 y 5.2 noticias, en media, entre las 20 primeras.

Día	6-5		7-5		8-5		9-5		10-5		Media	
Usuario	R+	R-	R+	R-	R+	R-	R+	R-	R+	R-	R+	R-
0	18	3	16	2	11	5	17	3	21	3	16.6	3.2
1	23	5	8	6	10	3	9	2	5	9	11.0	5.0
2	16	15	8	19	10	19	8	18	14	16	11.2	17.4
3	20	2	15	3	9	2	10	1	22	2	15.2	2.0
4	30	5	12	1	16	2	6	2	9	0	14.6	2.0
5	30	5	35	1	33	1	27	0	37	1	32.4	1.6
6	72	4	72	5	43	3	53	4	41	2	56.2	3.6
7	39	6	21	4	20	3	24	2	10	1	22.8	3.2
8	77	0	80	0	72	0	60	0	51	0	68.0	0.0
9	23	0	47	2	41	7	42	5	61	1	42.8	3.0
10	22	7	16	5	17	4	17	4	16	6	17.6	5.2
Media	33.6	4.7	30.0	4.4	25.6	4.5	24.8	3.7	26.1	3.7	28.0	4.2

Tabla 5.4. Número de noticias realimentadas cada día por cada usuario, de manera positiva o negativa (R+/R-).

5.5. Colección de evaluación 2.0

En este caso, la colección se obtuvo durante 14 días, del 1 de Diciembre de 2003 al 19 de Diciembre de 2003, excluyendo fines de semana y festivos, es decir, del 1 al 5, del 9 al 12 y del 15 al 19. El número de noticias bajadas cada día se refleja en la Tabla 5.5. Se utilizaron las siguientes 7 secciones: nacional, internacional, deportes, economía, sociedad, cultura y gente. No se pudo utilizar la sección de opinión porque pasó a ser de pago. El número total de noticias fue 1099.

Día	Noticias	Día	Noticias	Día	Noticias
1-12-2003	95	9-12-2003	76	15-12-2003	86
2-12-2003	75	10-12-2003	76	16-12-2003	81
3-12-2003	87	11-12-2003	85	17-12-2003	73
4-12-2003	71	12-12-2003	82	18-12-2003	72
5-12-2003	76			19-12-2003	64

Tabla 5.5. Número de noticias por día en la segunda colección de evaluación.

Esta colección [Díaz&Gervás04c] se utilizó en el sistema de personalización 2.0, en el cual se implementan los tres procesos de personalización, con el modelo de usuario completo, es decir, secciones, categorías y palabras clave.

El segundo paso de la construcción de la colección es obtener la definición de los intereses a largo plazo de varios modelos de usuario para poder empezar a ejecutar el sistema y obtener la primera personalización de contenidos.

El mecanismo utilizado para dar difusión al sistema fue el envío de correos electrónicos solicitando participación, a todos los profesores de la Ingeniería Técnica de Informática de Sistemas del CES Felipe II de Aranjuez, a todos los profesores del grupo de investigación Gaia de la Universidad Complutense de Madrid y a un profesor del departamento de Comunicación de la Universidad Rey Juan Carlos I de Madrid. Adicionalmente se solicitó la participación de los alumnos mediante explicación en clase y se permitió cualquier tipo de difusión adicional. Las carreras sobre las que se hizo especial explicación en clase fueron informática, periodismo, y publicidad y relaciones públicas.

En el mensaje de difusión se indicó a los posibles usuarios en que consistía el proceso de evaluación. En este caso, los usuarios recibían un mensaje cada día con la siguiente información asociada a cada noticia: título, autor, sección, resumen automático, enlace a la noticia completa e iconos de realimentación. Los usuarios debían emitir sus evaluaciones interaccionando con las noticias mediante la pulsación o no de los iconos de realimentación. Además los usuarios tenían que rellenar un cuestionario inicial de evaluación en el momento de registrarse y un cuestionario final al dejar de utilizar el sistema.

Se registraron inicialmente 104 usuarios, aunque sólo se consideran 102, puesto que 2 usuarios se registraron por error 2 veces. El primer día de funcionamiento del sistema se registraron otros 2 usuarios. Finalmente, el tercer día se registraron otros dos, dando un número total de 106 usuarios, a partir del tercer día. A partir del cuarto día empezaron a darse de baja usuarios, produciendo variaciones en el número de usuarios del sistema. El número de usuarios reales del sistema cada día se presenta en la Tabla 5.6.

Día	Usuarios	Día	Usuarios	Día	Usuarios
1-12-2003	102	9-12-2003	104	15-12-2003	101
2-12-2003	104	10-12-2003	102	16-12-2003	100
3-12-2003	106	11-12-2003	102	17-12-2003	100
4-12-2003	105	12-12-2003	101	18-12-2003	99
5-12-2003	105			19-12-2003	98

Tabla 5.6. Número de usuarios por día.

En cuanto al tipo de usuarios que realizaron la evaluación, se puede observar en la Tabla 5.7, que estuvo compuesto en un 72.6% por alumnos, frente a un 20.8% de profesores, el 6.6% fueron otros profesionales de distintos tipos. Dentro de los profesores, el grupo mayoritario fue el de profesores que imparten clase en ingeniería informática (16% del total), seguido por profesores de otras carreras o profesores de instituto (3.8% del total). También participó un profesor de periodismo. Dentro de los alumnos el grupo más numeroso fue el de periodismo (34% del total), seguido por alumnos de la licenciatura de publicidad y relaciones públicas (29.2% del total) y por último, alumnos de informática (9.4% del total). Del grupo de otros profesionales, 2 tenían relación con la informática y los otros 5 no tenían relación ni con la informática ni con el periodismo. Otra estadística obtenida fue que, curiosamente, el número de mujeres fue exactamente el mismo de hombres, es decir, 53. En cuanto al navegador utilizado el 88.9% utilizó Internet Explorer, el 6.7% Mozilla, el 2.2% Netscape y el 2.2% otro diferente.

	Número	Porcentaje
Profesor informática	17	16.0
Profesor periodismo	1	0.9
Profesor otros	4	3.8
Estudiante informática	10	9.4
Estudiante periodismo	36	34.0
Estudiante publicidad	31	29.2
Otros informática	2	1.9
otros	5	4.7
total	106	100

Tabla 5.7. Tipos de usuarios.

En conclusión, el grupo es suficientemente heterogéneo y numeroso como para incluir distintos tipos de comportamiento frente al sistema que permitan extraer conclusiones significativas de su funcionamiento.

Los modelos de usuario iniciales fueron construidos por los usuarios la semana anterior al comienzo del experimento. Esto permite que el sistema disponga de información sobre los modelos para poder realizar la personalización del primer día. La excepción la constituyen los 4 usuarios que se registraron en el sistema después del comienzo del experimento.

Estos perfiles iniciales, los de los 106 usuarios, contienen información sobre los intereses a largo plazo del usuario, esto es, secciones, categorías y palabras clave. Estos intereses son introducidos por el usuario manualmente utilizando una escala de 0 a 3.

	Secciones	Categorías	Palabras clave
Media	5.0	9.6	2.8
Máximo	7	14	18
Mínimo	0	0	0

Tabla 5.8. Estadísticas sobre el número de elementos seleccionados en los perfiles de usuario.

La Tabla 5.8 muestra el número de secciones, categorías y palabras clave elegidas por los usuarios en sus perfiles iniciales, esto es, en sus modelos a largo plazo. En general, los usua-

rios eligieron varias secciones, varias categorías y varias palabras clave. Estos modelos contenían, en media, 5 secciones, 9.6 categorías y 2.8 palabras clave. El hecho de que las secciones y las categorías se seleccionen marcando casillas mientras que las palabras clave tengan que ser introducidas por el usuario hace que éste último tipo de preferencia sea menos utilizado por los usuarios. Sin embargo, el número máximo de palabras clave seleccionadas por un usuario fue 18.

	Secciones	Categorías	Palabras clave
Valores seleccionados (porcentaje)	103 (97.2%)	102 (96.2%)	85 (80.2%)
Media seleccionados	5.1	10.0	3.5

Tabla 5.9. Estadísticas sobre el número de usuarios que eligieron cada uno de los métodos de selección.

Es interesante resaltar (Tabla 5.9) que 3 usuarios (2.8%) no seleccionaron ninguna sección, 4 (3.8%) no eligieron ninguna categoría y 21 (19.8%) no escogieron ninguna palabra clave. También es un dato significativo que 30 usuarios (28.3%) eligieran todas las secciones como relevantes y 28 (26.4%) eligieron todas las categorías. Por tanto, los métodos menos intuitivos se utilizaron menos. Los usuarios que eligieron secciones, seleccionaron 5.1 de media. Los que eligieron categorías, marcaron 10.0, y los que eligieron palabras clave, seleccionaron 3.5. Esto es, cuando un usuario optó por un método de selección, seleccionó más de 1 posibilidad.

Por último, hay que resaltar que hubo un usuario que introdujo un perfil vacío, es decir, no introdujo ninguna sección, ninguna categoría y ninguna palabra clave.

La Tabla 5.10 muestra los pesos asignados a cada una de las secciones. Se puede observar que la sección más importante para los usuarios es cultura, un 88.7% de los usuarios la eligieron, seguida de nacional con 82.1% e internacional con un 81.1%. Teniendo en cuenta sólo a los usuarios que asignaron pesos a cada sección, la sección de nacional tiene mayor promedio (0.82) en los valores de los pesos asignados por los usuarios, seguida de cultura con 0.77. La menos elegida fue deportes, 52.8% de los usuarios, seguida de economía, 57.5%. La de menor peso promedio es economía (0.53), seguida de gente, con 0.64.

	Nacional	Internacional	Economía	Sociedad	Cultura	Deportes	Gente
Media	0.68	0.58	0.30	0.51	0.68	0.39	0.39
Valores seleccionados	87 (82.1%)	86 (81.1%)	61 (57.5%)	81 (76.4%)	94 (88.7%)	56 (52.8%)	64 (60.4%)
Media seleccionados	0.82	0.71	0.53	0.67	0.77	0.73	0.64
Peso = 1	46 (43.4%)	26 (24.5%)	9 (8.5%)	26 (24.5%)	41 (38.7%)	24 (22.6%)	14 (13.2%)

Tabla 5.10. Pesos asignados a las distintas secciones por los usuarios.

También se puede resaltar que los usuarios variaron sus pesos entre las distintas secciones, es decir, no utilizaron sólo el criterio me interesa/no me interesa, lo cual demuestra que el sistema de asignación de pesos es atractivo para los usuarios. La sección que más fue elegida con peso 1 fue nacional, la seleccionaron el 43.4% de los usuarios, seguida de cultura (38.7%) y la que menos fue economía, sólo el 8.5%, seguida de gente, con el 13.2%.

	ArCu	CiTe	CiSo	DeOc	EcNe	EdFo	EsDi
Media	0.59	0.54	0.47	0.54	0.27	0.51	0.72
Valores seleccionados	82 (77.4%)	79 (74.5%)	72 (67.9%)	79 (74.5%)	56 (52.8%)	74 (69.8%)	93 (87.7%)
Media seleccionados	0.76	0.73	0.70	0.73	0.51	0.73	0.83
Peso = 1	33 (31.1%)	34 (32.1%)	24 (22.6%)	35 (33.0%)	8 (7.5%)	31 (29.2%)	53 (50.0%)

Tabla 5.11. Pesos asignados a las primeras 7 categorías por los usuarios.

Las Tablas 5.11 y 5.12 muestra los pesos asignados a cada una de las categorías. Las abreviaturas tienen los siguientes significados: ArCu, Arte y Cultura; CiTe, Ciencia y Tecnología; CiSo, Ciencias Sociales; DeOc, Deportes y Ocio; EcNe, Economía y Negocios; EdFo, Educación y Formación; EsDi, Espectáculos y Diversión; InOr, Internet y Ordenadores; MaCo, Materiales de Consulta; MeCo, Medios de Comunicación; PoGo, Política y Gobierno; Sa, Salud; So, Sociedad; ZoGe, Zonas Geográficas.

Se puede observar que la categoría más importantes para los usuarios es Espectáculos y Diversión, un 87.7% de los usuarios la eligieron, seguida de Medios de Comunicación, con un 80.2%. La categoría de Medios de Comunicación tiene mayor promedio (0.86) en los valores de los pesos asignados por los usuarios, seguida de Espectáculos y Diversión con 0.83. La menos elegida es Economía y Negocios, 52.8% de los usuarios, seguida de Zonas Geográficas, 53.8%. La de menor peso promedio es Economía y Negocios (0.51), seguida de Zonas Geográficas, con 0.63.

También se puede resaltar que los usuarios variaron sus pesos entre las distintas categorías, es decir, no utilizaron sólo el criterio me interesa/no me interesa, lo cual demuestra que el sistema de asignación de pesos es atractivo para los usuarios. La categoría que más fue elegida con peso 1 fue Medios de Comunicación, la seleccionaron el 51.9% de los usuarios, y la que menos fue Economía y Negocios, sólo el 7.5%.

	InOr	MaCo	MeCo	PoGo	Sa	So	ZoGe
Media	0.46	0.38	0.69	0.44	0.37	0.41	0.34
Valores seleccionados	73 (68.9%)	62 (58.5%)	85 (80.2%)	73 (68.9%)	67 (63.2%)	66 (62.3%)	57 (53.8%)
Media seleccionados	0.67	0.65	0.86	0.64	0.58	0.66	0.63
Peso = 1	24 (22.6%)	14 (13.2%)	55 (51.9%)	18 (17.0%)	10 (9.4%)	18 (17.0%)	9 (8.5%)

Tabla 5.12. Pesos asignados a las últimas 7 categorías por los usuarios.

En cuanto a los pesos de las palabras clave cabe destacar que el 80.2% de los usuarios introdujo alguna palabra clave. En cuanto a los pesos, el 73.6% de los usuarios asignó peso 1. En todo caso, el peso medio fue de 0.98 puesto que, de las 304 palabras clave introducidos por los usuarios, sólo 18 palabras clave fueron asignadas con peso distinto de 1 (a 16 palabras clave se les asignó 0.66 y a 2 palabras clave, 0.33), que corresponden a 9 (8.5%) usuarios distintos.

Hay que resaltar que de las 304 palabras clave introducidas por todos los usuarios 42 fueron nombres propios (39.6%) y casi todas los demás se referían a temas generales. Las más utilizadas fueron las siguientes: publicidad, por 19 usuarios; cine, por 14 y relaciones públicas, por 12. Tanto la primera como la última palabra fueron introducidas por los alumnos de publicidad y relaciones públicas, mientras que el término cine fue introducido por usuarios de distintos grupos. Otras palabras muy utilizadas fueron: Real Madrid, 8 usuarios; música, 7; periodismo, 6; deportes y ocio, 5; linux, comunicación, literatura y radio, 4; universidad, televisión, fútbol y tenis, 3.

Los nombres propios permiten una personalización muy concreta sobre temas de interés relacionados con personas o lugares que raramente dejaran de interesar al usuario, mientras que las palabras generales permiten enviar al usuario información que puede que no llegue a interesarle.

El tercer paso de la construcción de la colección es la obtención de los juicios de relevancia de cada uno de los usuarios respecto a todas las noticias, durante los 14 días. El proceso de obtención es similar al de la colección anterior, es decir, interacción del usuario con las noticias recibidas a través de los iconos de realimentación que representan realimentación positiva/realimentación negativa. En este caso, el resumen es el generado automáticamente por el sistema. También hay que tener en cuenta que, como los juicios de relevancia finales que se manejan son juicios binarios, se toman como relevantes sólo las noticias con realimentación positiva, mientras que se toman como no relevantes aquéllas con realimentación negativa o sobre las que no se haya efectuado ninguna realimentación.

Además se indica a los usuarios que sus juicios de realimentación no se deben basar sólo en su perfil a largo plazo sino que deben reflejar sus necesidades de información, precisamente los intereses que no sean acordes a su modelo a largo plazo deberían ser capturados por el modelo a corto plazo.

Hay que tener en cuenta que sólo se generan juicios de relevancia para aquellos usuarios que los introduzcan mediante las pulsaciones de los iconos de realimentación. Por tanto, no habrá tantos juicios como usuarios, sino que habrá tantos juicios como usuarios hayan emitido juicio sobre al menos una noticia. Además hay que resaltar que se produjo otra incidencia en el sistema: algunos correos fueron rebotados por los servidores de algunos usuarios, principalmente debido a saturaciones de los buzones. Evidentemente, esto imposibilito a los usuarios realimentar el sistema en esos casos.

Día	Usuarios	Rebot.	Día	Usuarios	Rebot.	Día	Usuarios	Rebot.
1-12-2003	50	11	9-12-2003	38	10	15-12-2003	35	13
2-12-2003	53	12	10-12-2003	35	13	16-12-2003	33	19
3-12-2003	38	14	11-12-2003	37	14	17-12-2003	25	16
4-12-2003	45	10	12-12-2003	31	10	18-12-2003	27	16
5-12-2003	44	11				19-12-2003	28	14

Tabla 5.13. Número de usuarios con juicios emitidos y mensajes rebotados, por día.

La Tabla 5.13 muestra el número de usuarios con juicios, cada día. Es patente que el interés de los usuarios por el sistema fue decreciendo a lo largo del tiempo, desde los 53 que juzgaron noticias el segundo día a los 25 del antepenúltimo día, con 37.1 como media. El esfuerzo para juzgar cada día aproximadamente 100 noticias no era trivial, lo que hace comprensible esta disminución. Estos juicios de usuario, son los que se utilizarán para evaluar el sistema. El número total de evaluaciones realizadas fue de 37854 (usuarios por día de Tabla 5.24 y noticias por día de Tabla 5.5).

En cuanto a la cantidad de mensajes rebotados, se mantuvo entre 10 y 19, con 13.1 de media. Tiene su explicación debido a que muchos de los usuarios eran alumnos con cuentas de correo en servidores gratuitos como hotmail, yahoo, wanadoo, etc., donde las cuotas asociadas no suelen ir más allá de 1 Mb y cada mensaje ocupaba alrededor de 200 Kb.

Num días	Usuarios	Mensajes rebotados
0	23	69
14	3	4
>=10	22	5
>=5 y <10	25	11
>=1 y <5	36	21
Media	4.9	1.7

Tabla 5.14. Número de días que unos usuarios emitieron juicios y otros no recibieron mensajes porque fueron rebotados.

La Tabla 5.14 muestra que la distribución de juicios de usuario y mensajes rebotados no fue siempre la misma durante todos los días. Hubo sólo 3 usuarios que emitieron juicios de relevancia todos los días, por 23 que no lo hicieron ningún día. De entre los 83 que realizaron algún juicio, 22 lo hicieron más de 10 días ó 10 días, 25 juzgaron entre 5 y 9 días, ambos inclusive, y 36 lo hicieron menos de 5 días. La media de juicios de usuario por día fue de 4.9.

Esto indica que muy pocos usuarios fueron fieles al sistema durante todos los días de funcionamiento del mismo, en el sentido de que emitieron sus juicios sobre las noticias. También se observa que hubo 23 usuarios que nunca emitieron juicios, aunque entre estos están algunos de los que no recibieron los mensajes porque su servidor de correo los rebotó, ya sea un día concreto o todos los días.

En cuanto a los mensajes rebotados, hubo 4 usuarios que no recibieron correo ningún día, mientras que hubo 69 que nunca rebotaron ningún correo. De entre los 37 que rebotaron algún correo, 5 lo hicieron más de 10 veces ó 10 veces, 11 rebotaron entre 5 y 9 correos, y 21 lo hicieron menos de 5 veces. La media de mensajes rebotados por usuario fue de 1.7 correos.

	Noticias	Media	Máximo	Mínimo
1-12-2003	95	35.9	95	1
2-12-2003	75	34.6	75	1
3-12-2003	87	34.8	87	1
4-12-2003	71	32.0	71	1
5-12-2003	76	32.8	76	1
9-12-2003	76	29.4	75	1
10-12-2003	76	29.4	76	1
11-12-2003	85	31.7	85	1
12-12-2003	82	36.8	82	1
15-12-2003	86	34.9	84	3
16-12-2003	81	32.3	81	1
17-12-2003	73	31.3	73	1
18-12-2003	72	30.0	72	1
19-12-2003	64	26.9	64	1
Media	78.5	32.3	78.3	1.1

Tabla 5.15. Estadísticas sobre el número de juicios de usuario por día y por usuario.

La Tabla 5.15 muestra las estadísticas sobre el número de noticias indicadas como relevantes por los usuarios durante los 14 días que duró el experimento. El número medio de juicios de relevancia por usuario, por día, varía entre 26.9 y 36.8, con 32.3 como valor medio. Los valores van disminuyendo según avanzan los días de funcionamiento del sistema por la misma razón que disminuye el número de usuarios, la tarea de juzgar 100 noticias por día no es trivial.

Los valores sufren una variación considerable entre distintos usuarios, hay usuarios que emiten juicios (relevantes o no relevantes) sobre muchas noticias, 95 de máximo, mientras que hay usuarios que se limitan a emitir un sólo juicio.

Se han considerado como juicios de usuario "válidos" para la evaluación del sistema aquellos que afectan, al menos, a 10 noticias. Esta consideración se basa en que el resto de los usuarios no han utilizado el sistema de manera consistente, sino que simplemente lo han usado por curiosidad y sin prestarle toda la atención necesaria como para que los juicios emitidos sean significativos. La distribución de usuarios por día con menos de 10 juicios es 12, 12, 8, 10, 9, 10, 9, 12, 8, 9, 8, 9, 10, 6. Esto supone eliminar una media de 9.4 usuarios por día, lo cual deja una media de 27.6 usuarios con juicios relevancia, por día.

A partir de ahora todos los valores que aparezcan estarán relacionados con estos juicios de usuario "válidos".

	Noticias	Media	Máximo	Mínimo
1-12-2003	95	45.7	95	11
2-12-2003	75	43.3	75	10
3-12-2003	87	42.9	87	10
4-12-2003	71	39.2	71	10
5-12-2003	76	40.3	76	11
9-12-2003	76	38	75	10
10-12-2003	76	38	76	10
11-12-2003	85	44.4	85	12
12-12-2003	82	47.7	82	13
15-12-2003	86	45.1	84	11
16-12-2003	81	41.5	81	12
17-12-2003	73	45.6	73	10
18-12-2003	72	45.1	72	11
19-12-2003	64	33.2	64	10
Media	78.5	42.1	78.3	10.8

Tabla 5.16. Estadísticas sobre el número de juicios de usuario por día y por usuario, para los usuarios "válidos".

La Tabla 5.16 muestra las estadísticas sobre el número de noticias indicadas como relevantes por los usuarios con más de 10 juicios por día. El número medio de noticias relevantes por día varía entre 33.2 y 47.7, con 42.1 como valor medio. Los valores van disminuyendo según avanzan los días de funcionamiento del sistema. El número total de evaluaciones realizadas por los usuarios válidos fue 30616.

Los valores sobre el número de juicios "válidos" (Tabla 5.17) sufren variación entre distintos usuarios. Hay una media de 11.7 usuarios (42.4%) que emiten juicios (relevantes o no relevantes) sobre muchas noticias (>= 50 noticias). El siguiente intervalo de juicios con más usuarios es entre 10 y 20 noticias. La media es de 6.4 usuarios (23%). El resto de intervalos de 10 noticias tienen un número de usuarios similar: 4.1 (14.7%), 3.2 (11.6%) y 3.1 (11.4%). de media.

El cuarto paso de la construcción de la colección es la obtención de los juicios sobre la realimentación del usuario sobre cada una de las noticias, durante los 14 días. Para ello se utiliza el mismo mecanismo que en la colección anterior, correspondencia con las pulsaciones en los iconos de realimentación en base a 3 posibilidades: realimentación positiva, no realimentación o realimentación negativa.

	>10 y <20	>=20 y <30	>=30 y <40	>=40 y <50	>= 50	>=10
1-12-2003	9	8	2	2	19	38
2-12-2003	6	6	7	6	18	41
3-12-2003	8	4	1	3	14	30
4-12-2003	6	5	8	5	12	35
5-12-2003	9	4	5	5	13	35
9-12-2003	9	3	3	4	9	28
10-12-2003	7	6	0	5	9	26
11-12-2003	4	7	1	2	11	25
12-12-2003	3	3	1	3	14	23
15-12-2003	6	2	4	2	12	26
16-12-2003	7	3	3	3	9	25
17-12-2003	4	1	1	2	9	16
18-12-2003	3	1	4	1	8	17
19-12-2003	8	4	5	1	7	22
Media	6.4	4.1	3.2	3.1	11.7	27.6

Tabla 5.17. Estadísticas sobre el número de usuarios "válidos" según el número de juicios por día.

	Noticias	Media R+	Media R-	Media R+ y R-
1-12-2003	95	23.0	23.3	45.7
2-12-2003	75	21.5	22.3	43.3
3-12-2003	87	20.5	23.1	42.9
4-12-2003	71	19.9	19.3	39.2
5-12-2003	76	21.7	19.2	40.3
9-12-2003	76	17.7	21.1	38.0
10-12-2003	76	19.1	19.9	38.0
11-12-2003	85	22.0	23.4	44.4
12-12-2003	82	22.3	26.4	47.7
15-12-2003	86	21.0	25.0	45.1
16-12-2003	81	20.9	21.5	41.5
17-12-2003	73	17.5	28.2	45.6
18-12-2003	72	18.6	29.9	45.1
19-12-2003	64	17.5	17.3	33.2
Media	78.5	20.2	22.9	42.1

Tabla 5.18. Estadísticas sobre el número de noticias realimentadas cada día, positiva o negativamente (R+/R-), por usuario y por día.

La Tabla 5.18 muestra el número de noticias realimentadas, tanto positiva como negativamente (R+/R-), en media para todos los usuarios con juicios "válidos", durante cada uno de los 14 días que duró el experimento. La media de noticias realimentadas positivamente se sitúa en torno a 20 noticias por usuario y por día, mientras que la media de negativas se sitúa en torno a 23 noticias por usuario y por día. Es decir que se efectúa realimentación positiva sobre un 25.8% de noticias, en media, y realimentación negativa, sobre un 29.1% de las mismas. Por tanto los usuarios tienden a realimentar negativamente un poco más que positivamente. La media con respecto al total de noticias realimentadas por usuario y por día es de 42.1 (53.7% de las noticias).

	R+ > R-	R- > R+
1-12-2003	25	15
2-12-2003	28	15
3-12-2003	16	14
4-12-2003	23	13
5-12-2003	24	12
9-12-2003	16	12
10-12-2003	16	11
11-12-2003	16	9
12-12-2003	12	12
15-12-2003	13	13
16-12-2003	18	7
17-12-2003	6	11
18-12-2003	8	9
19-12-2003	13	9
media	16.7	11.6

Tabla 5.19. Estadísticas sobre el número de usuarios que realimentan más noticias positiva que negativamente y viceversa, cada día.

También se ha observado que hay 10 usuarios que sólo realimentan positivamente y 1 que sólo realimenta negativamente, el resto realimentan tanto positiva como negativamente. En general, hay más usuarios que realimentan más noticias positiva que negativamente, 16.7 frente a 11.6 de media por día (Tabla 5.19). Por lo tanto, se puede concluir que los usuarios, en media, realimentan más positiva que negativamente, aunque el número de noticias realimentadas negativamente es ligeramente mayor.

5.6. Resumen y conclusiones del capítulo

En este capítulo se han presentado las distintas colecciones de evaluación desarrolladas a lo largo de la tesis para evaluar distintas versiones de sistemas de personalización de contenidos. Todas ellas utilizan el periódico ABC como fuente de información y contienen información sobre los perfiles de usuario, los juicios de relevancia y los juicios de realimentación de los usuarios.

Las dos primeras minicolecciones se utilizaron para evaluar, de manera preliminar, la selección de contenidos y se basan en los datos de un sólo día de evaluación. Además los juicios de relevancia son generados por un experto humano y no por los propios usuarios. La tercera minicolección es un primer intento de evaluación de la presentación de resultados.

La primera colección se utilizó con el sistema de personalización 1.0. Almacena información para juzgar los tres procesos de personalización: aproximadamente 100 noticias por día (8 secciones), 11 modelos de usuario (sin categorías), juicios de relevancia y juicios de realimentación. Además se almacena esta información para los 5 días (del 6 al 10 de Mayo de 2002) que dura el proceso de evaluación. Por otro lado, los juicios recogidos son los de los propios usuarios.

Tras realizar la colección anterior se hizo patente la necesidad de evaluaciones de más usuarios y durante más días para poder obtener resultados más significativos. Esto llevó a otra colección más completa durante más días.

La segunda colección es similar a la anterior pero contiene 106 usuarios durante 14 días (del 1 al 19 de Diciembre de 2003), con aproximadamente 80 noticias por día. En este caso, el modelo de usuario utilizado es el modelo completo (secciones, categorías y palabras clave). Además se utilizaron sólo 7 secciones, porque una de las 8 que se usaba en la versión anterior pasó a ser de pago. Los juicios también son los de los propios usuarios.

Capítulo 6
EXPERIMENTOS REALIZADOS

6.1. Introducción

En este capítulo se van a explicar y evaluar las distintas propuestas de personalización de contenidos web aplicadas. Estas propuestas de sistemas han ido evolucionando a lo largo del tiempo según se han ido conociendo el efecto de las técnicas elegidas a través de la evaluación efectuada. En realidad es un proceso paralelo a la evolución producida en las colecciones de evaluación descritas en el tema anterior.

En el apartado 6.2 se describirán los experimentos preliminares [Gervás *et al.* 99; Díaz&Gervás00; Díaz *et al.* 00a; Díaz *et al.* 00b; Díaz *et al.* 01b; Buenaga *et al.* 01; Acero *et al.* 01] realizados en torno a la selección de contenidos con las dos primeras minicolecciones de evaluación y en relación con la presentación de resultados efectuada con la tercera minicolección de colección.

Estos primeros experimentos van a dar lugar a evaluaciones iniciales de la propuesta de personalización presentada en esta tesis. Las ideas recogidas de esas evaluaciones serán utilizadas en los dos sistemas de personalización propuestos, donde se realizará una evaluación mucho más completa a través de la aplicación de la metodología de evaluación planteada en el Capítulo 4.

En el apartado 6.3 se describirá el sistema de personalización 1.0 [Acero&Alcojor01; Díaz01], en el cual se simplificará el sistema de selección, eliminando las categorías, pero se introducirá el proceso de adaptación. Se realizará una evaluación utilizando la colección de evaluación 1.0 [Díaz&Gervás03; Díaz&Gervás04a; Díaz&Gervás04b].

El sistema de personalización 2.0 será descrito en el apartado 6.4, en él se aplicarán técnicas similares a las utilizadas en el sistema anterior, introduciendo las categorías en el proceso de selección. Además se utilizará una colección de evaluación mucho más completa, con más usuarios y más días, la colección de evaluación 2.0 [Díaz&Gervás04c; Díaz *et al.* 05a; Díaz *et al.* 05b].

Por último, en el apartado 6.5 se mostrará un resumen y las conclusiones del capítulo.

6.2. Experimentos preliminares

En este apartado se van a comentar los resultados obtenidos en varios experimentos preliminares realizados sobre los procesos de selección de contenidos y presentación de resultados. Estos experimentos utilizaron las minicolecciones de evaluación descritas en el apartado 5.2.

6.2.1. Primer experimento preliminar

En el primer experimento preliminar, descrito en [Díaz *et al.* 00a], se trataba de evaluar la selección de contenidos. Los resultados son generados por el proceso de presentación, pero, en este caso, el resumen de cada noticia se obtiene a partir del resumen proporcionado por el editor, cuando éste está disponible y en caso contrario se utilizan las primeras líneas de la noticia. Además el usuario puede elegir cuántas noticias iba a contener el mensaje que iba a recibir.

El modelo de usuario utilizado incluye los 3 métodos de selección: secciones, categorías y palabras clave. La representación de las categorías se realizó utilizando únicamente la información contenida en las páginas de primer nivel de Yahoo! España. Por otro lado, el usuario podía determinar el peso que asignaba a cada uno de los métodos de selección. Sin embargo, una carencia de este sistema es que no se normalizan los resultados provenientes de los distintos sistemas de referencia antes de ser combinados.

Esta primera evaluación se realizó utilizando la primera minicolección preliminar descrita en el capítulo anterior: 44 usuarios durante 11 días, del 28 de Enero al 10 de Febrero del 2000; 8 secciones con aproximadamente 100 noticias por día.

Se hicieron dos tipos de evaluaciones: una evaluación cualitativa, basada en impresiones de los usuarios recogidas en formularios (Apéndice I), y una evaluación cuantitativa, obtenida a partir de juicios de las noticias relevantes y no relevantes enviadas por el sistema. Estos juicios de relevancia son obtenidos por un experto humano para un día concreto. A partir de estas evaluaciones se pueden obtener dos resultados de recall y precisión: uno basado en las impresiones de los usuarios y otro basado en los juicios de relevancia extraídos por el experto humano.

Comparando los resultados obtenidos en la evaluación cualitativa respecto a la evaluación cuantitativa se pudo observar que la precisión es alto en ambos casos pero el recall es mucho más bajo en la evaluación cuantitativa. La razón es que un usuario considera una noticia como relevante si se refiere a algo que es relevante para él, independientemente de si pertenece a una sección, a una categoría o si contiene una palabra clave. Por otro lado, el recall bajo es debido al límite máximo en el número de noticias recibidas fijado por los usuarios: con un modelo de usuario con unas pocas secciones y unas pocas categorías el número de noticias relevantes es demasiado alto para ser capturado en un nivel de recall máximo fijado por los usuarios mediante el número de noticias recibidas (14.1 como valor medio).

Otra justificación de la diferencia en los dos tipos de evaluación es que los juicios del experto humano están basados en un OR-lógico de todos los intereses reflejados en el perfil, mientras que los usuarios tienden a buscar más un AND-lógico, sino de todos los intereses, si de grupos de intereses que constituyan un tipo de necesidad de información.

El experimento mostró resultados bastante prometedores mediante la utilización de la personalización de contenidos basada en un modelo de usuario que se basa en la combinación de tres sistemas diferentes de clasificación: secciones, categorías y palabras clave. Sin embargo, se detectaron algunos problemas de entendimiento por parte de los usuarios del funcionamiento del sistema. Por lo tanto, es necesaria más ayuda, sobre todo para que los usuarios no tengan ninguna dificultad a la hora de rellenar sus perfiles de usuario. De hecho, este problema es el que llevó a la generación de 14 perfiles vacíos, secciones elegidas como sistema de clasificación, pero peso general de secciones igual a cero.

La evaluación se realizó con un único valor de recall y precisión, el correspondiente al límite máximo de noticias recibidas por el usuario. Se podría mejorar la evaluación si se tuvieran en cuenta distintos niveles para evaluar estas medidas, además las medias obtenidas en

realidad se están aplicando a distintos niveles de recall, lo cual puede dar lugar a interpretaciones poco claras.

Por otro lado, la evaluación cualitativa mostró resultados bastante aceptables, con una media de medio-alto para las distintas valoraciones efectuadas, lo cual confirma la utilidad del sistema para los usuarios.

Finalmente resaltar que el problema de la normalización de las relevancias obtenidas a partir de cada sistema de clasificación hizo que la relevancia proveniente de las secciones fuera mucho más significativa que la proveniente de categorías y palabras clave. Sin embargo, este problema no afectó a los resultados de evaluación porque los niveles de recall utilizados para obtener los valores de precisión incluían, en la mayoría de los casos, noticias provenientes de secciones seleccionadas por el usuario.

6.2.2. Segundo experimento preliminar

Este segundo experimento se realizó utilizando la segunda minicolección de evaluación descrita en el capítulo anterior: 66 modelos durante 1 día, el 9 de Febrero del 2000. Ese día hubo 109 noticias. Este experimento aparece descrito en [Díaz *et al.* 01a].

En este experimento se repitió la evaluación cuantitativa realizada en el primer experimento pero teniendo en cuenta 4 niveles de recall en lugar de uno. En la evaluación anterior el recall y la precisión se medían para un mismo nivel, el fijado por el usuario con su límite máximo de noticias. Sin embargo, esta evaluación es mejorable porque mezcla valores de recall y precisión a distintos niveles y porque sólo muestra un único valor. Además se crearon perfiles de distintos tipo para permitir interpretar mejor las contribuciones asociadas a cada uno de los sistemas de clasificación: secciones, categorías y palabras clave.

El estudio de los diferentes comportamientos de los tres métodos usados para especificar los intereses de los usuarios (secciones, categorías y palabras clave) mostró que la interacción entre los distintos factores que afectan al sistema no es trivial. Estos factores son el número de noticias por sección, el número de noticias por categoría, el número máximo de noticias recibidas fijado por el usuario, la relevancia de los contenidos para un día concreto y para un usuario concreto. Debido a que estos parámetros cambian diariamente en el caso de los servicios digitales de noticias, y que sus valores para un día concreto son independientes de los valores que tuvieran en días anteriores y posteriores, hacen imposible afirmar que uno de los métodos es siempre mejor, basándose en la evaluación de un día concreto. En particular, para los experimentos realizados los mejores resultados se obtuvieron con las secciones, recall y precisión altos, y los peores con las categorías, recall y precisión bajos.

Por otro lado, aparecieron problemas no triviales intrínsecos al funcionamiento de este tipo de sistemas: como un periódico debe venir lleno de noticias cada día, haya ocurrido algo interesante o no, las secciones de los periódicos tienen que contener una serie de noticias cada día independientemente de su relevancia con el nombre de la sección, y por lo tanto un mensaje personalizado que se base sólo en secciones puede contener noticias que no están relacionadas con los intereses del usuario correspondiente.

El problema de la no-normalización de las distintas contribuciones de los distintos métodos de clasificación hace esta comparación más difícil puesto que las relevancias provenientes de las secciones son siempre superiores a las de los otros sistemas de referencia. Además no se utiliza realmente el potencial de la combinación de métodos.

Por otro lado, la representación de las categorías es mejorable porque algunas de ellas se han representado con muy pocas palabras, lo cual no establece una representación significa-

tiva del contenido de la categoría que se pueda utilizar para clasificar noticias correctamente. El problema se agudiza debido a que los resultados no se normalizaron cuando se combinaron los distintos métodos de clasificación.

6.2.3 . Tercer experimento preliminar

El tercer experimento preliminar se centró en el proceso de presentación de resultados en forma de resúmenes. Se realizó una evaluación de este sistema con la tercera minicolección de evaluación descrita en el capítulo anterior: 3 usuarios durante 1 día, el 9 de Febrero del 2000. Hubo 109 noticias ese día. Este experimento aparece descrito en [Acero *et al.* 01].

En este caso, el modelo de usuario manejado incluía los 3 métodos de selección: secciones, categorías y palabras clave. Además se normalizaron tanto las contribuciones de los distintos sistemas de referencia en la selección de contenidos, como las contribuciones de las distintas heurísticas en la generación de resúmenes, y el usuario podía determinar el peso que asignaba a cada uno de los métodos de selección.

Se generaron distintos tipos de resúmenes para los 3 usuarios para poder compararlos entre sí, y con las noticias completas, y establecer cuál era la mejor opción. En este caso se generaron, por un lado, un resumen genérico para cada una de las 109 noticias del día de la evaluación, y por otro, 5 tipos de resúmenes genérico-personalizados diferentes para cada una de las 109 noticias y para cada uno de los 3 usuarios. Es decir, se generaron 16 resúmenes distintos para una misma noticia. En total, 1744 resúmenes diferentes.

En general, los resultados fueron muy alentadores aunque las diferencias fueron poco significativas. Es aconsejable realizar una evaluación con un mayor número de usuarios y noticias, para comprobar que es mejor la utilización de resúmenes genérico-personalizados frente a resúmenes genéricos en sistemas de personalización de noticias. Además se debería incorporar la posibilidad de generar resúmenes personalizados sólo con las heurísticas de personalización para evaluar su efectividad. Por otro lado, la generación de resúmenes constituidos por las primeras frases de las noticias sería una buena línea base con la que comparar los resultados.

6.3. Sistema de personalización de noticias 1.0

En el primer sistema de personalización se utilizaba selección, adaptación y presentación de resultados [Acero&Alcojor01; Díaz01]. El sistema seleccionaba de entre todas las noticias de un día, aquellas que eran más relevantes para cada usuario según su perfil y éstas eran enviadas en un correo electrónico resumidas automáticamente. El usuario podía realimentar el sistema mediante sus opiniones lo cual afectaba tanto a la selección como a la presentación de los siguientes días.

Un cambio importante fue la eliminación de las categorías como sistema de referencia en el modelo a largo plazo del usuario. La razón de esta supresión estuvo basada en el objetivo de construir un sistema donde los modelos de usuario fueran más fácilmente interpretables, permitiendo obtener resultados más claros de la evaluación de los distintos procesos de personalización. Además las peores evaluaciones obtenidas en los experimentos preliminares provenían de las categorías. Por lo tanto, en este sistema el modelo de usuario maneja sólo 2 métodos de selección: secciones y palabras clave.

En este sistema se normalizan cuando son combinadas, tanto las contribuciones de los distintos sistemas de referencia en la selección de contenidos, como los provenientes de las heurísticas de construcción de resúmenes en la presentación de resultados.

Por otro lado, fue eliminada la posibilidad de que los usuarios asignaran un peso general a cada uno de los métodos de clasificación, tomando estos valores como variables en los experimentos. Esto permite identificar cuáles son las contribuciones de cada método de clasificación para poder determinar cuál es la mejor combinación posible de los mismos.

La evaluación de este sistema implicó la evaluación de los tres procesos de personalización. Se realizó una evaluación de este sistema utilizando la colección de evaluación 1.0, descrita en el capítulo anterior, donde los juicios de relevancia son emitidos por los propios usuarios [Díaz&Gervás03; Díaz&Gervás04a; Díaz&Gervás04b].

6.3.1 . Selección de contenidos

La ecuación utilizada para realizar la selección es la (3.8) con $\beta=0$ (no se utilizan las categorías). Los valores de α y χ serán utilizados como variables en los experimentos y tendrán valor 1 cuando sea considerada la posibilidad de que el valor correspondiente sea distinto de cero.

Se van a tratar de demostrar las hipótesis referidas a la selección de contenidos a través del experimento 1 planteado en el apartado 4.4.2.1. En este caso, todos las combinaciones que incluyen a las categorías no han sido tenidas en cuenta.

6.3.1.1. Experimento 1. Combinación de secciones y palabras clave, dentro del modelo a largo plazo.

Se han calculado los valores de recall y precisión normalizados para cada uno de las combinaciones de mecanismos de selección posibles, esto es, sólo secciones (S), sólo palabras clave (P) y combinación de secciones y palabras clave (SP). Estos experimentos se han repetido durante los 5 días de evaluación. En total, para cada mecanismo de selección se han realizado 5709 evaluaciones.

6.3.1.2. Resultados

Las medias de recall y precisión normalizados, por usuario, sobre todos los días y en media se presentan en la Tabla 6.1. Estos resultados muestran que la combinación de secciones y palabras clave da mejor resultado que la utilización de cada una de ellas por separado, tanto en términos de recall normalizado (nR) como en términos de precisión normalizada (nP).

Como resultado adicional, se observa que la selección es mejor cuando sólo se utilizan las secciones que cuando sólo se utilizan las palabras clave, tanto en recall como en precisión normalizados.

	6-Mayo		7-Mayo		8-Mayo		9-Mayo		10-Mayo		Medias	
	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP
SP	0.616	0.507	0.569	0.436	0.627	0.510	0.630	0.501	0.642	0.508	0.617	0.493
S	0.595	0.444	0.566	0.410	0.610	0.461	0.625	0.457	0.619	0.444	0.603	0.443
P	0.550	0.395	0.505	0.325	0.547	0.376	0.533	0.383	0.553	0.378	0.538	0.371

Tabla 6.1. Recall y precisión normalizados para las distintas combinaciones de secciones y palabras clave, para cada día y en media.

Los datos medios confirman los resultados obtenidos usando el test de significancia estadística sobre los diferentes mecanismos de selección para los distintos usuarios durante todos los días de la evaluación. Los porcentajes de mejora sobre los valores medios de recall y precisión normalizados se resumen en la Tabla 6.2. La combinación de secciones y palabras clave mejora la precisión respecto a las secciones en un 11.2% y la mejora respecto a las palabras clave en un 32.7%. Las secciones dan mejores resultados en precisión (19.3%) que las palabras clave. Las mejoras en recall mantienen la tendencia aunque los porcentajes son menores. Todos los resultados son estadísticamente significativos.

	SP > S	SP > P	S > P
% nP	11.2	32.7	19.3
% nR	2.3	14.7	12.2

Tabla 6.2. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de recall y precisión normalizados.

Por lo tanto, se confirma la primera hipótesis (H1): la precisión en la selección de las noticias con respecto a los modelos de usuario, usando sólo el modelo a largo plazo, es mejor si se utiliza una combinación de todos los sistemas de referencia que si se utiliza cualquier otra combinación.

Una vez establecida la combinación de secciones y palabras clave como la mejor opción para identificar los intereses a largo plazo, este mecanismo se empleará, mientras no se diga lo contrario, para todos los demás experimentos presentados en este sistema.

6.3.2. Adaptación del modelo de usuario

En el proceso de adaptación lo que se evalúa es el efecto en la selección de contenidos debido a la adaptación del modelo de usuario. La ecuación utilizada para realizar esta selección es la (3.10) con $\beta=0$ (no se utilizan las categorías). Los valores de α, χ y ε serán utilizados como variables en los experimentos y tendrán valor 1 cuando sea considerada la posibilidad de que el valor correspondiente sea distinto de cero.

Se van a tratar de demostrar las hipótesis referidas a la adaptación de contenidos a través de los experimentos 2 y 3 planteados en los apartados 4.5.2.1 y 4.5.2.2. En este caso, todas las combinaciones que incluyen a las categorías no han sido tenidas en cuenta.

6.3.2.1. Experimento 2. Combinación de modelos a corto y largo plazo.

Se han calculado los valores de recall y precisión normalizados para cada uno de los mecanismos de selección indicados, esto es, largo plazo con secciones (L(S)), largo plazo con palabras clave (L(P)), largo plazo con combinación de secciones y palabras clave (L(SP)), corto plazo (O), largo plazo con secciones y corto plazo (L(S)O), largo plazo con palabras clave y corto plazo (L(P)O), largo plazo con combinación de secciones y palabras clave, y corto plazo (L(SP)O). Estos experimentos se han repetido durante los últimos 4 días de evaluación ya que el primer día no hay modelo a corto plazo. En total, para cada mecanismo de selección se han realizado 4301 evaluaciones.

6.3.2.2. Resultados

Las medias globales de recall y precisión normalizados por usuario sobre todos los días se presentan en la Tabla 6.3. Estos resultados muestran que la combinación de los modelos a largo y corto plazo dan mejor resultado que la utilización de cada una de ellos por separado, tanto en términos de precisión normalizada como en términos de recall normalizado.

Como resultado adicional, se puede observar que los modelos a largo plazo ofrecen mejores resultados que el modelo a corto plazo, excepto para el caso en el que se utilizan únicamente las palabras clave.

	7-Mayo		8-Mayo		9-Mayo		10-Mayo		Medias	
	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP
L(SP)O	0.577	0.483	0.619	0.540	0.640	0.555	0.659	0.544	0.624	0.531
L(SP)	0.569	0.436	0.627	0.510	0.630	0.501	0.642	0.508	0.617	0.489
L(S)O	0.566	0.477	0.596	0.529	0.616	0.536	0.640	0.520	0.604	0.515
L(S)	0.566	0.410	0.610	0.461	0.625	0.457	0.619	0.444	0.605	0.443
L(P)O	0.510	0.413	0.589	0.486	0.562	0.483	0.598	0.472	0.565	0.464
L(P)	0.505	0.325	0.547	0.376	0.533	0.383	0.553	0.378	0.534	0.365
O	0.523	0.422	0.553	0.457	0.498	0.367	0.471	0.350	0.511	0.399

Tabla 6.3. Recall y precisión normalizados para las distintas combinaciones de largo y corto plazo, para cada día y en media.

Los porcentajes de mejora sobre los valores medios de precisión normalizados se resumen en la Tabla 6.4.

Respecto a las palabras clave, el único resultado estadísticamente significativo que se obtiene es que es mejor combinar el largo y el corto plazo que utilizar sólo el largo (26.9%). También hay mejora relativa de la combinación respecto al modelo a corto plazo (16.1%), pero no es estadísticamente significativa. El modelo a corto plazo supera al modelo a largo plazo (8.5%), pero no significativamente.

En cuanto a las secciones, los resultados estadísticamente significativos son que la combinación es siempre mejor que los modelos por separado (16.3% y 29.0%). El modelo a largo plazo supera al modelo a corto plazo, pero no significativamente (10.9%).

En cuanto a la combinación de secciones y palabras clave, todos los resultados son estadísticamente significativos, es decir, la combinación supera a ambos modelos en solitario (8.5% y 32.9%) y el largo plazo es mejor que el corto (22.4%).

Palabras clave	% nP	Secciones	% nP	Combinación	% nP
L(P)O > L(P)	26.9	L(S)O > L(S)	16.3	L(SP)O > L(SP)	8.5
L(P)O > O	16.1	L(S)O > O	29.0	L(SP)O > O	32.9
O > L(P)	8.5	L(S) > O	10.9	L(SP) > O	22.4

Tabla 6.4. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de precisión normalizada.

Las mejoras en recall (Tabla 6.5) mantienen la tendencia aunque los porcentajes son menores. Se observa una pequeña diferencia, en recall, a favor de las combinaciones de largo y corto plazo respecto a la utilización de largo plazo en solitario, es decir, que a pesar de no ser estadísticamente mejor de manera significativa, si es mejor ligeramente, si tenemos en cuenta los valores medios de recall normalizado. Las combinaciones superan al corto plazo, y el largo supera al corto siempre de manera significativa, excepto en el caso de las palabras clave.

Palabras clave	% nR	Secciones	% nR	Combinación	% nR
L(P)O > L(P)	5.8	L(S) > L(S)O	0.1	L(SP)O > L(SP)	1.1
L(P)O > O	10.6	L(S)O > O	18.2	L(SP)O > O	22.0
L(P) > O	4.5	L(S) > O	18.3	L(SP) > O	20.7

Tabla 6.5. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de recall normalizado.

Por lo tanto, podemos concluir que se confirma la segunda hipótesis (H2): la selección de las noticias con respecto a los modelos de usuario, es mejor si se utiliza una combinación de los modelos a corto y largo plazo, que si alguno de ellos se utiliza en solitario. Como resultado adicional, se observa que, en la mayoría de los casos, el modelo a largo plazo en solitario es mejor que el modelo a corto plazo en solitario.

6.3.2.3. Experimento 3. Combinación de largo y corto plazo, combinando secciones y palabras clave para el modelo a largo plazo.

Se han calculado los valores de recall y precisión normalizados para cada uno de los mecanismos de selección con combinación de los modelos a largo y corto plazo, esto es, largo plazo con secciones y corto plazo (L(S)O), largo plazo con palabras clave y corto plazo (L(P)O), largo plazo con combinación de secciones y palabras clave, y corto plazo (L(SP)O). Estos experimentos se han repetido durante los últimos 4 días de evaluación ya que el primer día no hay modelo a corto plazo. En total, para cada mecanismo de selección se han realizado 4301 evaluaciones.

6.3.2.4. Resultados

Los resultados mostrados en la Tabla 6.6 muestran que la combinación de los modelos a largo y corto plazo dan mejor resultado si se utiliza para el modelo a largo plazo la combinación de secciones y palabras clave que si sólo se utilizan secciones o palabras clave por separado, tanto en términos de precisión normalizada como en términos de recall normalizado.

También se observa que la combinación utilizando las secciones para el modelo a largo plazo es mejor que la combinación utilizando las palabras clave.

	7-Mayo		8-Mayo		9-Mayo		10-Mayo		Medias	
	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP
L(SP)O	0.577	0.483	0.619	0.540	0.640	0.555	0.659	0.544	0.624	0.531
L(S)O	0.566	0.477	0.596	0.529	0.616	0.536	0.640	0.520	0.604	0.515
L(P)O	0.510	0.413	0.589	0.486	0.562	0.483	0.598	0.472	0.565	0.464

Tabla 6.6. Recall y precisión normalizados para las distintas combinaciones de secciones y palabras clave del modelo a largo, cuando se utiliza la combinación de largo y corto plazo, para cada día y en media.

Los porcentajes de mejora sobre los valores medios de recall y precisión normalizados se resumen en la Tabla 6.7. La combinación de largo y corto plazo utilizando secciones y palabras clave mejora la precisión, de manera no significativa, respecto a la utilización de sólo secciones en un 2.9%, y la mejora significativamente respecto a sólo palabras clave en un 12.6%. La opción de sólo secciones da mejores resultados significativos en precisión (11.2%) que la de sólo palabras clave.

En términos de recall normalizado, la combinación de largo y corto plazo utilizando la combinación secciones y palabras clave es significativamente mejor que cuando se utiliza sólo palabras clave (9.4%) y que cuando se utiliza sólo secciones (3.1%). No hay mejora significativa entre sólo secciones y sólo palabras clave, aunque si hay ligera mejoría (6.9%).

	% nP	% nR
L(SP)O > L(S)O	2.9	3.1
L(SP)O > L(P)O	12.6	9.4
L(S)O > L(P)O	11.2	6.9

Tabla 6.7. Porcentajes de mejora respecto a los valores medios de recall y precisión normalizados, para las distintas combinaciones de secciones y palabras clave del modelo a largo, cuando se utiliza la combinación de largo y corto plazo.

Esto confirma la tercera hipótesis (H3): la selección de las noticias con respecto a los modelos de usuario, utilizando una combinación de los modelos a corto y largo plazo, es mejor si se utiliza para el modelo a largo plazo la combinación de todos los sistemas de referencia que si se utiliza cualquier otra combinación. Como resultado adicional se puede observar que la combinación de ambos modelos es mejor si se utilizan sólo las secciones para el largo plazo y que si sólo se utilizan las palabras clave.

Habiendo establecido la combinación de los modelos a largo plazo, con secciones y palabras clave, y corto plazo, como la mejor opción, este mecanismo será utilizado, mientras no se diga lo contrario, para el resto de los experimentos presentados en este sistema.

6.3.3. Presentación de resultados

En este sistema el modelo de usuario maneja 2 métodos de selección: secciones y palabras clave. Además el usuario no puede asignar un peso a los distintos métodos de clasificación sino que estos son tomados como variables en los experimentos a realizar.

Las ecuaciones utilizadas para generar los resúmenes son la (3.15) y la (3.16). Por lo tanto, se utilizan las palabras clave del modelo a largo plazo y los términos de realimentación del modelo a corto plazo. Los valores de φ, γ, η y $\varkappa$ serán utilizados como variables en los experimentos y tendrán valor 1 cuando sea considerada la posibilidad de que el valor correspondiente sea distinto de cero.

Por otro lado, la selección de los distintos tipos de resúmenes se realiza con la ecuación (3.10) con las variables α, χ y ε igual a 1, y por otro lado, β igual a 0, al no usar las categorías.

Se van a tratar de demostrar las hipótesis referidas a la presentación de resultados a través de los experimentos 4 y 5 planteados en los apartados 4.6.2.1 y 4.6.2.2.

6.3.3.1. Experimento 4. Generación de resúmenes personalizados.

Varias colecciones de evaluación diferentes se han generado para cada usuario, cada una de ellas consiste en el conjunto de resúmenes de las noticias originales obtenidos mediante la aplicación de cada uno de los métodos de generación de resúmenes personalizados indicados, es decir, habrá una colección para cada usuario de resúmenes personalizados usando el modelo a corto plazo (Rp(O)), otra equivalente con el modelo a largo plazo (Rp(L)) y una tercera usando la combinación de ambos modelos (Rp(LO)). Es decir, se han generado 3 resúmenes distintos para una misma noticia.

Ya que la colección de resúmenes es diferente para cada usuario el proceso de evaluación debe hacerse usuario por usuario, es decir, en realidad se realizan tantos procesos de evaluación por cada tipo de resumen como usuarios diferentes haya en el sistema.

En cada caso, se han calculado los valores de recall y precisión normalizados. Estos experimentos se han repetido durante los últimos 4 días de evaluación, ya que el primer día no hay corto plazo. En total se han generado 12903 resúmenes personalizados diferentes (11 usuarios, 391 noticias distintas, 3 tipos de resúmenes por usuario y por noticia).

6.3.3.2. Resultados

Los resultados mostrados en la Tabla 6.8 muestran que la combinación de los modelos a largo y corto plazo para la generación de resúmenes personalizados da mejor resultado que la utilización de cada uno de ellos por separado en términos de precisión y recall normalizados. Entre la utilización de los modelos en solitario es mejor la opción de utilizar sólo el corto plazo.

	6-Mayo		7-Mayo		8-Mayo		9-Mayo		10-Mayo		Medias	
	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP
Rp(LO)	0.616	0.508	0.581	0.486	0.620	0.538	0.632	0.535	0.657	0.550	0.621	0.523
Rp(O)	0.610	0.497	0.588	0.493	0.622	0.550	0.624	0.526	0.642	0.524	0.617	0.518
Rp(L)	0.616	0.508	0.574	0.482	0.611	0.532	0.625	0.523	0.631	0.507	0.611	0.510

Tabla 6.8. Recall y precisión normalizados para las distintas combinaciones de los modelos a largo plazo y corto plazo para la generación de resúmenes personalizados.

Los porcentajes de mejora sobre los valores medios de recall y precisión normalizados se resumen en la Tabla 6.9. La combinación de largo y corto plazo mejora la precisión significativamente respecto al largo plazo en un 2.5% y la mejora, pero no significativamente, respec-

to al corto plazo en un 1.0%. El corto plazo da mejores resultados en precisión (1.4%) que el largo plazo, de manera significativa. Las mejoras en recall mantienen la tendencia aunque los porcentajes son menores y las diferencias no son significativas.

	% nP	% nR
Rp(LO) > Rp(L)	2.5	1.6
Rp(LO) > Rp(O)	1.0	0.7
Rp(O) > Rp(L)	1.4	0.9

Tabla 6.9. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de recall y precisión normalizados.

Aunque la diferencia entre la combinación de modelos y el modelo a corto plazo no es estadísticamente significativa, si es ligeramente mejor, lo cual nos permite retener la cuarta hipótesis (H4) como una hipótesis razonable: los resúmenes, obtenidos usando sólo la heurística de personalización, son mejores si se utiliza una combinación de los modelos a largo y corto plazo, que si alguno de ellos se utiliza en solitario. Como resultado adicional, se observa que el modelo a corto plazo en solitario es mejor que el modelo a largo plazo en términos de precisión normalizada.

De aquí en adelante, cuando nos refiramos a resúmenes personalizados (Rp), estaremos refiriéndonos a la personalización obtenida mediante la combinación de los modelos a largo y corto plazo.

6.3.3.3. Experimento 5. Combinación de heurísticas para la generación de resúmenes.

Las colecciones de resúmenes base (Rb) y de resúmenes genéricos (Rg) serán las mismas para todos los usuarios, puesto que no dependen de ningún aspecto del modelo de usuario, es decir, no están personalizadas. Esto permitirá evaluar a todos los usuarios a la vez.

Sin embargo, habrá una colección diferente por usuario de resúmenes personalizados (Rp) y de resúmenes genérico-personalizados (Rgp), puesto que éstas sí dependen del modelo del usuario. Habrá que evaluar por separado a cada usuario con respecto a su colección de resúmenes personalizados, por un lado, y respecto a su colección de resúmenes genérico-personalizados, por otro. Se han generado 24 resúmenes distintos para una misma noticia.

En cada caso, se han calculado los valores de recall y precisión normalizados. Estos experimentos se han repetido durante los 5 días de evaluación. En total se han generado 12456 resúmenes diferentes (24 resúmenes por noticia, 519 noticias). Un ejemplo de resúmenes generados aparece en el Apéndice III.

6.3.3.4. Resultados

Los resultados mostrados en la Tabla 6.10 muestran que los resúmenes personalizados ofrecen mejores resultados con respecto a la precisión normalizada de la información seleccionada que los resúmenes genéricos y que los resúmenes genérico-personalizados. Los resúmenes genérico-personalizados son mejores que los resúmenes genéricos, y los resúmenes genéricos son mejores que los resúmenes base.

También se puede observar en la misma tabla que los resúmenes personalizados son peores que las noticias completas (N) bajo el mismo criterio. Aunque esta diferencia no es esta-

dísticamente significativa si se observa una ligera mejora de la noticia completa respecto al resumen personalizado.

Los resúmenes personalizados son mejores que los resúmenes basados en las primeras frases de las noticias, con una mejora estadísticamente significativa. En términos de recall normalizado, las tendencias se mantienen.

	6-Mayo		7-Mayo		8-Mayo		9-Mayo		10-Mayo		Medias	
	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP	nR	nP
N	0.616	0.507	0.577	0.483	0.619	0.540	0.640	0.555	0.659	0.544	0.622	0.526
Rp	0.616	0.508	0.581	0.486	0.620	0.538	0.632	0.535	0.657	0.550	0.621	0.523
Rgp	0.612	0.497	0.563	0.448	0.603	0.527	0.625	0.521	0.649	0.539	0.611	0.506
Rg	0.607	0.479	0.563	0.452	0.612	0.526	0.631	0.529	0.652	0.538	0.613	0.505
Rb	0.607	0.479	0.557	0.450	0.598	0.513	0.622	0.522	0.649	0.536	0.607	0.500

Tabla 6.10. Recall y precisión normalizados para los distintos tipos de resúmenes.

Los porcentajes de mejora sobre los valores medios de recall y precisión normalizados se resumen en la Tabla 6.11. Los resúmenes personalizados mejoran la precisión respecto a los resúmenes genérico-personalizados en un 3.3%, la mejoran respecto a los resúmenes genéricos en un 3.7% y la mejoran respecto a los resúmenes base en un 4.7%. En todos los casos la mejora es estadísticamente significativa.

Las noticias completas dan mejor resultado que los resúmenes personalizados, pero sólo hay una diferencia del 0.5%. Por último, los resúmenes genérico-personalizados mejoran los resúmenes genéricos en un ligero 0.3% y los resúmenes genéricos ofrecen mejores resultados que los resúmenes base en un 1.0%. Todos estos resultados no son estadísticamente significativos.

Las mejoras en recall mantienen la tendencia aunque los porcentajes son menores. Se observa una anomalía en la diferencia en recall entre resúmenes genérico-personalizados y resúmenes genéricos, esto es debido a la pequeña diferencia entre ambos tipos de resúmenes, que de hecho, no sólo no es significativa sino que hay igual número de casos a favor de los resúmenes genérico-personalizados que a favor de los resúmenes genéricos. El resto de los casos tampoco son estadísticamente significativos excepto la comparación entre resúmenes personalizados y resúmenes base.

	Rp > Rgp	Rp > Rg	Rp > Rb	N > Rp	Rgp > Rg	Rg > Rb
% nR	1.8	1.4	2.4	0.2	-0.4	1.1
% nP	3.3	3.7	4.7	0.5	0.3	1.0

Tabla 6.11. Porcentajes de mejora de los distintos tipos de resúmenes respecto a los valores medios de recall y precisión normalizados.

Estos resultados nos permiten confirmar la hipótesis H5: Los resúmenes obtenidos usando sólo la heurística de personalización son mejores, con respecto a la información seleccionada por los usuarios, que los resúmenes obtenidos extrayendo las primeras frases del texto del documento completo.

También nos permiten retener, aunque en términos de recall normalizado los resultados no sean significativos, la hipótesis H6: Los resúmenes obtenidos utilizando únicamente la heurística de personalización son mejores con respecto a la información seleccionada por los

usuarios, que los resúmenes obtenidos usando las heurísticas de generación de resúmenes genéricos y que los resúmenes obtenidos usando una combinación de heurísticas.

Por último, también queda confirmada parcialmente la hipótesis H7: Los resúmenes obtenidos usando sólo la heurística de personalización son peores que las noticias completas con respecto a la información seleccionada por los usuarios.

6.3.4. Conclusiones del sistema de personalización 1.0

El sistema de personalización 1.0 se ha evaluado respecto a los 3 procesos de personalización mediante la utilización de la colección 1.0. El modelo de usuario sólo contiene secciones y palabras clave.

La evaluación del proceso de selección de contenidos ha mostrado que la combinación de secciones y palabras clave ofrece mejores resultados, en términos de recall y precisión normalizados, que la utilización de secciones o palabras clave por separado. Por otro lado, se ha observado que las secciones ofrecen mejores resultados que las palabras clave.

De la evaluación del proceso de adaptación del modelo de usuario se ha obtenido que la combinación de largo y corto plazo es mejor que la utilización de los modelos en solitario tanto en términos de precisión normalizada como en términos de recall normalizado. El modelo a largo plazo en solitario es mejor que el modelo a corto plazo en solitario tanto en términos de recall como en términos de precisión.

También se ha obtenido que los mejores resultados en la combinación de los modelos a largo y corto plazo se producen cuando se combinan secciones y palabras clave en el modelo a largo plazo, como cabría esperar de los resultados obtenidos en los experimentos sobre la selección de contenidos. La opción de sólo secciones es mejor que sólo palabras clave de manera significativa en términos de precisión normalizada pero no en términos de recall normalizado.

La evaluación de la presentación de resultados permite concluir que los resúmenes personalizados utilizando una combinación de los modelos a largo y corto plazo es mejor que los otros tipos de resúmenes en términos de precisión normalizada, pero no en términos de recall. Por otro lado, las noticias completas ofrecen una ligera mejora no significativa frente a los resúmenes personalizados, lo cual quiere decir que la pérdida de información para el usuario es muy pequeña con este tipo de resúmenes.

Se puede extraer de los datos obtenidos que los resúmenes genéricos ofrecen resultados muy similares a los obtenidos con los resúmenes base, es decir, con las primeras frases de los documentos. Esto parece indicar que la heurística de posición está obteniendo valores mucho más altos que la heurística de palabras significativas. En cualquier caso, aunque una propuesta basada en resúmenes obtenidos a partir de las primeras frases de un documento puede dar buenos resultados en términos de resúmenes indicativos generales, no da tan buenos resultados cuando lo que se busca son resúmenes personalizados para cada usuario, donde es crucial retener en el texto aquellos fragmentos que estén relacionados con el perfil del usuario.

Esto explica porque los resúmenes genérico-personalizados dan tan poco rendimiento a pesar de ser una combinación de buenas técnicas: dado un límite fijo en la longitud de un resumen, la inclusión de sentencias seleccionadas por las heurísticas genéricas en la mayoría de los casos empuja hacia el final del resumen a aquellas frases que habrían sido útiles desde el punto de vista de la personalización.

6.4. Sistema de personalización 2.0

En el segundo sistema de personalización se utilizaba selección, adaptación y presentación de resultados [Díaz&Gervás04c; Díaz *et al.* 05a; Díaz *et al.* 05b]. El sistema seleccionaba de entre todas las noticias de un día, aquellas que eran más relevantes para cada usuario según su perfil y éstas eran enviadas en un correo electrónico resumidas automáticamente. El usuario podía realimentar el sistema mediante sus opiniones lo cual afectaba tanto a la selección como a la presentación de los siguientes días.

En este sistema se volvieron a añadir las categorías como sistema de referencia para el modelo a largo plazo, pero mejorando la representación de las mismas mediante la utilización de la información de las categorías de segundo nivel. El número medio de palabras por categoría es aproximadamente de 250, por 72 del primer experimento.

Por otro lado, al igual que en el sistema anterior, las contribuciones son normalizadas cuando son combinadas y los pesos generales de los métodos de clasificación no son determinados por los usuarios sino que son utilizados como variables en los experimentos

La evaluación de este sistema implicó la evaluación de los tres procesos de personalización. Se realizó una evaluación de este sistema utilizando la colección de evaluación 2.0, donde los juicios de relevancia son emitidos por los propios usuarios.

6.4.1. Selección de contenidos

La ecuación utilizada para realizar la selección es la (3.8). Los valores de α, β y χ serán utilizados como variables en los experimentos y tendrán valor 1 cuando sea considerada la posibilidad de que el valor correspondiente sea distinto de cero.

Se van a tratar de demostrar las hipótesis referidas a la selección de contenidos a través del experimento 1 planteado en el apartado 4.4.2.1. En este caso, se han tenido en cuenta todas las combinaciones.

6.4.1.1. Experimento 1. Combinación de secciones, categorías y palabras clave, dentro del modelo a largo plazo.

Se han calculado los valores de recall y precisión normalizados para cada uno de las combinaciones de mecanismos de selección posibles, esto es, sólo secciones (S), sólo categorías (C), sólo palabras clave (P), combinación de secciones y categorías (SC), combinación de secciones y palabras clave (SP), combinación de categorías y palabras clave (CP) y combinación de secciones, categorías y palabras clave (SCP). Estos experimentos se han repetido durante los 14 días de evaluación. En total, para cada mecanismo de selección se han realizado 30616 evaluaciones.

6.4.1.2. Resultados

Las medias globales de recall y precisión normalizados, por usuario y por día, se presentan en la Tabla 6.12. Estos resultados muestran que la combinación de secciones, categorías y palabras clave dan mejor resultado que cualquier otra combinación de ellas, tanto en términos de recall normalizado como en términos de precisión normalizada.

Como resultado adicional, se observa que la combinación de sistemas es la mejor opción y dentro de las opciones individuales la que ofrece mejores resultados son las categorías se-

guidas de las secciones. En particular la combinación de categorías y secciones es la segunda mejor combinación, seguida de la combinación de categorías con palabras clave. La peor elección son las palabras clave, seguida por las secciones.

	SCP	SC	CP	SP	C	S	P
nP	0.588	0.579	0.513	0.505	0.496	0.451	0.349
nR	0.685	0.680	0.617	0.650	0.605	0.638	0.530

Tabla 6.12. Recall y precisión normalizados medios para las distintas combinaciones de secciones, categorías y palabras clave dentro del modelo a largo plazo.

Los datos medios confirman los resultados obtenidos usando el test de significancia estadística sobre los diferentes mecanismos de selección para los distintos usuarios durante todos los días de la evaluación. Los porcentajes de mejora sobre los valores medios de precisión normalizada se resumen en la Tabla 6.13. La combinación de secciones, categorías y palabras clave mejora la precisión respecto a la combinación de secciones y categorías (1.6%), pero no significativamente. Tampoco hay diferencia significativa entre categorías y palabras clave frente a secciones y palabras clave (1.6%), ni entre secciones y palabras clave frente a categorías (1.8%). Sí hay mejora significativa entre secciones y categorías frente a categorías y palabras clave (12.9%), también entre categorías frente a secciones (10.0%), y entre secciones frente a palabras clave (29.2%). El resto de comparaciones entre sistemas de clasificación según el orden mostrado para la precisión en la Tabla 6.13 es estadísticamente significativo.

La utilización de las categorías es la que ofrece mayor potencial en la clasificación, seguida de las secciones. Este potencial hace que la combinación de los 3 sistemas ofrezca los mejores resultados, seguido por los que combinan con las categorías. Además también ofrecen los mejores resultados cuando se usa un único método clasificación.

	SCP > SC	SC > CP	CP > SP	SP > C	C > S	S > P
% nP	1.6	12.9	1.6	1.8	10.0	29.2
	SCP > SC	SC > SP	SP > S	S > CP	CP > C	C > P
% nR	0.8	4.7	1.8	3.4	1.9	14.2

Tabla 6.13. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de recall y precisión normalizados.

Por otro lado, la combinación de secciones, categorías y palabras clave mejora el recall respecto a secciones y categorías (0.8%), pero no significativamente. Todas las demás comparaciones son estadísticamente significativas: secciones y categorías frente a secciones y palabras clave (4.7%), secciones y palabras clave frente a secciones (1.8%), secciones frente a categorías y palabras clave (3.4%), categorías y palabras clave frente a categorías (1.9%), y categorías frente a palabras clave (14.2%). El resto de comparaciones entre sistemas de clasificación según el orden mostrado para el recall en la Tabla 6.13 es estadísticamente significativo.

Para el recall es más interesante utilizar las secciones frente a las categorías. La combinación de los 3 sistemas de los mejores resultados, seguido por los que combinan con las secciones, seguido de las secciones y después las otras combinaciones. Esta diferencia entre recall y precisión es debido a las distintas formas que tienen de medir la efectividad estas métricas.

Por lo tanto, se confirma la primera hipótesis (H1): la selección de las noticias con respecto a los modelos de usuario, usando sólo el modelo a largo plazo, es mejor si se utiliza una combinación de todos los sistemas de referencia que si se utiliza cualquier otra combinación.

Una vez establecida la combinación de secciones, categorías y palabras clave como la mejor opción para identificar los intereses a largo plazo, este mecanismo se empleará, mientras no se diga lo contrario, para todos los demás experimentos presentados en este sistema.

6.4.2. Adaptación del modelo de usuario

El proceso de adaptación evalúa el efecto en la selección de contenidos debido a la adaptación del modelo de usuario. La ecuación utilizada para realizar esta selección es la (3.10). Los valores de α, β, χ y ε serán utilizados como variables en los experimentos y tendrán valor 1 cuando sea considerada la posibilidad de que el valor correspondiente sea distinto de cero.

Se van a tratar de demostrar las hipótesis referidas a la adaptación de contenidos a través de los experimentos 2 y 3 planteados en los apartados 4.5.2.1 y 4.5.2.2.

6.4.2.1. Experimento 2. Combinación de modelos a corto y largo plazo.

Se han calculado los valores de recall y precisión normalizados para cada uno de los mecanismos de adaptación posibles, esto es, largo plazo con secciones (L(S)), largo plazo con categorías (L(C)), largo plazo con palabras clave (L(P)), largo plazo con combinación de secciones y palabras clave (L(SP)), largo plazo con combinación de categorías y palabras clave (L(CP)), largo plazo con combinación de secciones y categorías (L(SC)), largo plazo con combinación de secciones, categorías y palabras clave (L(SCP)), corto plazo (O), largo plazo con secciones y corto plazo (L(S)O), largo plazo con categorías y corto plazo (L(C)O), largo plazo con palabras clave y corto plazo (L(P)O), largo plazo con combinación de secciones y palabras clave, y corto plazo (L(SP)O), largo plazo con combinación de secciones y categorías, y corto plazo (L(SC)O), largo plazo con combinación de categorías y palabras clave, y corto plazo (L(CP)O), largo plazo con combinación de secciones, categorías y palabras clave, y corto plazo (L(SCP)O). Estos experimentos se han repetido durante los últimos 13 días de evaluación ya que el primer día no hay modelo a corto plazo. En total, para cada mecanismo de selección se han realizado 29232 evaluaciones.

6.4.2.2. Resultados

Las medias globales de recall y precisión normalizados, por usuario y por día, se presentan en la Tabla 6.14. Estos resultados muestran que la combinación de los modelos a largo y corto plazo dan mejor resultado que la utilización de cada una de ellos por separado en términos de precisión normalizada, sea cual sea el sistema de clasificación elegido para el largo plazo. En términos de recall normalizado se repite la misma situación.

Como resultado adicional, se puede observar que los modelos a largo plazo ofrecen mejores resultados que el modelo a corto plazo, excepto para el caso en el que se utilizan únicamente las palabras clave.

	L(SCP)O	L(SC)O	L(SP)O	L(CP)O	L(S)O	L(C)O	L(P)O
nP	0.600	0.583	0.568	0.539	0.535	0.514	0.475
nR	0.691	0.681	0.669	0.633	0.652	0.614	0.583
	L(SCP)	L(SC)	L(SP)	L(CP)	L(S)	L(C)	L(P)
nP	0.583	0.574	0.503	0.509	0.448	0.492	0.349
nR	0.683	0.678	0.649	0.613	0.637	0.601	0.530
	O	O	O	O	O	O	O
nP	0.421	0.421	0.421	0.421	0.421	0.421	0.421
nR	0.545	0.545	0.545	0.545	0.545	0.545	0.545

Tabla 6.14. Precisión y recall normalizados para las distintas combinaciones de los modelos a largo plazo y corto plazo.

Para entender los malos resultados asociados a la opción de sólo corto plazo hay que indicar que la adaptación del sistema, es decir, la obtención de los términos de realimentación se realiza sobre las primeras noticias en el ranking del usuario. Los malos resultados surgen porque el primer día, al no haber criterios a largo plazo que ordenen las noticias, la presentación de las noticias al usuario es aleatoria, por tanto, la probabilidad de que las noticias relevantes para el usuario aparezcan las primeras es pequeño.

Los porcentajes de mejora sobre los valores medios de precisión normalizada se resumen en la Tabla 6.15.

La combinación de largo y corto plazo mejora siempre, de manera significativa, la precisión respecto al largo plazo, en solitario. Estas diferencias varían, siendo la menor de ellas la producida cuando se utiliza como largo plazo las secciones y las categorías (1.7%), seguida por secciones, categorías y palabras clave (3.0%). La mayor diferencia aparece cuando se utilizan sólo las palabras clave (36.2%), seguida por las secciones en solitario (19.4%).

Los sistemas de clasificación que dan mejores resultados en la selección son los que menos mejoran, con respecto al largo plazo, cuando se incluye el corto plazo y viceversa, los que dan peores resultados en la selección son los que más mejoran.

La combinación de largo y corto plazo también mejora siempre, de manera significativa, la precisión respecto al corto plazo, en solitario. Estas diferencias son siempre mayores que las diferencias entre largo y corto plazo y sólo largo plazo. El incremento en precisión varía, siendo el mayor de ellos el producido cuando se utiliza como largo plazo las secciones, las categorías y las palabras clave (42.4%), seguida por secciones y categorías (38.4%). El menor incremento aparece cuando se utilizan sólo las palabras clave (12.8%), seguida por las categorías en solitario (22.0%).

LC > L	% nP	LC > C	% nP	L > C	% nP
L(SCP)O > L(SCP)	3.0	L(SCP)O > O	42.4	L(SCP) > O	38.3
L(SC)O > L(SC)	1.7	L(SC)O > O	38.4	L(SC) > O	36.2
L(SP)O > L(SP)	13.1	L(SP)O > O	34.8	L(SP) > O	19.2
L(CP)O > L(CP)	5.9	L(CP)O > O	23.2	L(CP) > O	20.8
L(S)O > L(S)	19.4	L(S)O > O	26.9	L(S) > O	6.2
L(C)O > L(C)	4.5	L(C)O > O	22.0	L(C) > O	16.8
L(P)O > L(P)	36.2	L(P)O > O	12.8	O > L(P)	20.8

Tabla 6.15. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de precisión normalizados.

El largo plazo en solitario mejora, de manera significativa, la precisión respecto al corto plazo, en solitario, excepto en el caso de palabras clave para el largo plazo, en el cual es mejor significativamente el corto que el largo plazo. Estas diferencias son menores que las diferencias entre la combinación de modelos y sólo corto plazo. El incremento en precisión varía, siendo el mayor de ellos el producido cuando se utiliza como largo plazo las secciones, las categorías y las palabras clave (38.3%), seguida por secciones y categorías (36.2%). El menor incremento aparece cuando se utilizan sólo las secciones (6.2%), seguida por las categorías en solitario (16.8%). La diferencia entre sólo corto plazo y sólo largo con palabras clave es del 20.8%.

De nuevo, dan mayores incrementos las combinaciones de sistemas y peores los que incluyen sistemas en solitario. En todo caso, hay que resaltar que el corto plazo sea mejor que el largo sólo con palabras clave, esto es debido, como ya se observó en la selección a la mala precisión obtenida cuando se utilizan las palabras clave en solitario.

LC > L	% nR	LC > C	% nR	L > C	% nR
L(SCP)O > L(SCP)	1.2	L(SCP)O > O	26.9	L(SCP) > O	25.4
L(SC)O > L(SC)	0.5	L(SC)O > O	25.0	L(SC) > O	24.5
L(SP)O > L(SP)	3.2	L(SP)O > O	22.9	L(SP) > O	19.1
L(CP)O > L(CP)	3.4	L(CP)O > O	14.5	L(CP) > O	12.5
L(S)O > L(S)	2.3	L(S)O > O	19.7	L(S) > O	17.0
L(C)O > L(C)	2.1	L(C)O > O	12.7	L(C) > O	10.4
L(P)O > L(P)	10.0	L(P)O > O	7.1	O > L(P)	2.7

Tabla 6.16. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de recall normalizado.

Los porcentajes de mejora sobre los valores medios de recall normalizado se resumen en la Tabla 6.16.

La combinación de largo y corto plazo mejora el recall, de manera significativa, respecto al largo plazo en solitario en todos los casos, excepto en las combinaciones de los 3 sistemas de clasificación y la combinación de secciones y categorías, en los cuales la mejora no es estadísticamente significativa. Las diferencias varían, siendo la menor de ellas la producida cuando se utiliza como largo plazo las secciones y las categorías (0.5%), seguida por seccio-

nes, categorías y palabras clave (1.2%). La mayor diferencia aparece cuando se utilizan sólo las palabras clave (10.0%), seguida por categorías y palabras clave (3.4%).

En general, los sistemas de clasificación que dan mejores resultados en la selección son los que menos mejoran, con respecto al largo plazo, cuando se incluye el corto plazo y, viceversa, los que dan peores resultados en la selección son los que más mejoran. Los incrementos en recall son menores que los incrementos en precisión debido al comportamiento distinto de ambas métricas.

El largo plazo en solitario mejora siempre, de manera significativa, el recall respecto al corto plazo en solitario. Estas diferencias son siempre mayores que las diferencias entre largo y corto plazo y sólo largo plazo. El incremento en precisión varía, siendo el mayor de ellos el producido cuando se utiliza como largo plazo las secciones, las categorías y las palabras clave (26.9%), seguida por secciones y categorías (25.0%). El menor incremento aparece cuando se utilizan sólo las palabras clave (7.1%), seguida por las categorías en solitario (12.7%).

La combinación de sistemas de clasificación que dan mejores incrementos respecto al corto plazo son los formados por combinaciones de sistemas, siendo los peores los que incluyen sistemas en solitario.

El largo plazo en solitario mejora, de manera significativa, el recall respecto al corto plazo en solitario, excepto en el caso de palabras clave para el largo plazo, en el cual es mejor significativamente el corto que el largo plazo. El incremento en precisión varía, siendo el mayor de ellos el producido cuando se utiliza como largo plazo las secciones, las categorías y las palabras clave (25.4%), seguida por secciones y categorías (24.5%). El menor incremento aparece cuando se utilizan sólo las categorías (10.4%), seguida por las categorías más palabras clave (12.5%).

De nuevo, dan mayores incrementos las combinaciones de sistemas y peores los que incluyen sistemas en solitario. En todo caso, hay que resaltar que el corto plazo sea mejor que el largo sólo con palabras clave.

Por lo tanto, podemos concluir que se confirma la segunda hipótesis (H2): la selección de las noticias con respecto a los modelos de usuario, es mejor si se utiliza una combinación de los modelos a corto y largo plazo, que si alguno de ellos se utiliza en solitario. Como resultado adicional, se observa que, en la mayoría de los casos, el modelo a largo plazo en solitario es mejor que el modelo a corto plazo en solitario. Como resultado adicional, se observa que, en la mayoría de los casos, el modelo a largo plazo en solitario es mejor que el modelo a corto plazo en solitario.

6.4.2.3. Experimento 3. Combinación de largo y corto plazo, combinando secciones, categorías y palabras clave para el modelo a largo plazo.

Se han calculado los valores de recall y precisión normalizados para cada uno de los mecanismos de selección con combinación de los modelos a largo y corto plazo, esto es, largo plazo con secciones y corto plazo (L(S)O), largo plazo con palabras clave y corto plazo (L(P)O), largo plazo con combinación de secciones y palabras clave, y corto plazo (L(SP)O), largo plazo con combinación de secciones y categorías, y corto plazo (L(SC)O), largo plazo con combinación de categorías y palabras clave, y corto plazo (L(CP)O), largo plazo con combinación de secciones, categorías y palabras clave, y corto plazo (L(SCP)O). Estos experimentos se han repetido durante los últimos 13 días de evaluación ya que el primer día no hay modelo a corto plazo. En total, para cada mecanismo de selección se han realizado 29232 evaluaciones.

6.4.2.4. Resultados

Los resultados mostrados en la Tabla 6.17 muestran que en la combinación de largo y corto plazo, la combinación de secciones, categorías y palabras clave para el largo plazo, da mejor resultado que la utilización de cualquier otra combinación de sistemas de clasificación, en términos de precisión normalizada. Como resultado adicional, se observa que las opciones que presentan la combinación de varios sistemas en el largo plazo mejoran la precisión frente a los que no la presentan.

	L(SCP)O	L(SC)O	L(SP)O	L(CP)O	L(S)O	L(C)O	L(P)O
nP	0.600	0.583	0.568	0.539	0.535	0.514	0.475
nR	0.691	0.681	0.669	0.633	0.652	0.614	0.583

Tabla 6.17. Precisión y recall normalizados para las mejores combinaciones de los modelos a largo plazo y corto plazo.

Los porcentajes de mejora sobre los valores medios de recall y precisión normalizados se resumen en la Tabla 6.18. La combinación de secciones y categorías, más las palabras clave mejora significativamente la precisión respecto a secciones y categorías (2.9%). También hay mejora significativa entre secciones y categorías frente a secciones y palabras clave (2.7%), también entre secciones y palabras clave frente a categorías y palabras clave (5.3%) , y entre categorías frente a palabras clave (8.2%). Por otro lado, no hay diferencia significativa entre categorías y palabras clave frente a secciones (0.9%), ni entre secciones frente a categorías (4.0%). El resto de comparaciones entre sistemas de clasificación según el orden mostrado para la precisión en la Tabla 6.18 es estadísticamente significativo.

La combinación de sistemas de clasificación para el largo plazo que, combinados con el corto plazo, dan mejores resultados, son los formados por combinaciones de sistemas, siendo los peores los que incluyen sistemas en solitario.

	% nP		% nR
L(SCP)O > L(SC)O	2.9	L(SCP)O > L(SC)O	1.5
L(SC)O > L(SP)O	2.7	L(SC)O > L(SP)O	1.7
L(SP)O > L(CP)O	5.3	L(SP)O > L(S)O	2.7
L(CP)O > L(S)O	0.9	L(S)O > L(CP)O	2.9
L(S)O > L(C)O	4.0	L(CP)O > L(C)O	3.2
L(C)O > L(P)O	8.2	L(C)O > L(P)O	5.2

Tabla 6.18. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de recall y precisión normalizados.

La combinación de secciones, categorías y palabras clave mejora significativamente el recall respecto a secciones y categorías (1.5%). También hay mejora significativa entre secciones y categorías frente a secciones y palabras clave (1.7%), también entre secciones y palabras clave frente a secciones (2.7%), igualmente entre categorías y palabras clave frente a categorías (3.2%), y entre categorías frente a palabras clave (5.2%). Por otro lado, no hay diferencia significativa entre secciones frente a categorías y palabras clave (2.9%). El resto de comparaciones entre sistemas de clasificación según el orden mostrado para el recall en la Tabla 6.18 es estadísticamente significativo.

Para el recall, lo más interesante es utilizar las secciones frente a las categorías. La combinación de los 3 sistemas de los mejores resultados, seguido por los que combinan con las secciones, seguido de las secciones y después las otras combinaciones.

Esto confirma la tercera hipótesis (H3): la selección de las noticias con respecto a los modelos de usuario, utilizando una combinación de los modelos a corto y largo plazo, es mejor si se utiliza para el modelo a largo plazo la combinación de todos los sistemas de referencia que si se utiliza cualquier otra combinación.

Habiendo establecido la combinación de los modelos a largo y corto plazo, utilizando secciones, categorías y palabras clave, como la mejor opción, este mecanismo será utilizado, mientras no se diga lo contrario, para el resto de los experimentos presentados en esta tesis.

6.4.3. Presentación de resultados

En este sistema el modelo de usuario maneja los 3 métodos de selección: secciones, categorías y palabras clave. Además el usuario no puede asignar un peso a los distintos métodos de clasificación sino que estos son tomados como variables en los experimentos a realizar.

Las ecuaciones utilizadas para generar los resúmenes son la (3.15) y la (3.16). Por lo tanto, para generar los resúmenes se utilizan las palabras claves del modelo a largo plazo y los términos de realimentación del modelo a corto plazo. Los valores de φ, γ, η y $\varkappa$ serán utilizados como variables en los experimentos y tendrán valor 1 cuando sea considerada la posibilidad de que el valor correspondiente sea distinto de cero.

Por otro lado, la selección de los distintos tipos de resúmenes se realiza con la ecuación (3.10) con las variables α, β, χ y ε igual a 1.

Se van a tratar de demostrar las hipótesis referidas a la presentación de resultados a través de los experimentos 4 y 5 planteados en los apartados 4.6.2.1 y 4.6.2.2.

6.4.3.1. Experimento 4. Generación de resúmenes personalizados.

Varias colecciones de evaluación diferentes se han generado para cada usuario, cada una de ellas consiste en el conjunto de resúmenes de las noticias originales obtenidos mediante la aplicación de cada uno de los métodos de generación de resúmenes personalizados indicados, es decir, habrá una colección para cada usuario de resúmenes personalizados usando el modelo a corto plazo (Rp(O)), otra equivalente con el modelo a largo plazo (Rp(L)) y una tercera usando la combinación de ambos modelos (Rp(LO)). Se han generado 3 resúmenes distintos para una misma noticia para un mismo usuario.

Ya que la colección de resúmenes es diferente para cada usuario el proceso de evaluación debe hacerse usuario por usuario, es decir, en realidad se realizan tantos procesos de evaluación por cada tipo de resumen como usuarios diferentes haya en el sistema.

En cada caso, se han calculado los valores de recall y precisión normalizados. Estos experimentos se han repetido durante los últimos 13 días de evaluación, ya que el primer día no hay corto plazo.

En total se han generado 77598 resúmenes personalizados diferentes (28 usuarios de media por día, 1004 noticias distintas, 3 resúmenes distintos por usuario y por noticia).

6.4.3.2. Resultados

Los resultados medios por usuario y por día, mostrados en la Tabla 6.19, muestran que la combinación de los modelos a largo y corto plazo para la generación de resúmenes personalizados da mejor resultado que la utilización de cada uno de ellos por separado en términos de precisión y recall normalizados. Entre la utilización de los modelos en solitario es mejor la opción de utilizar sólo el corto plazo.

	nP	nR
Rp(LO)	0.592	0.684
Rp(O)	0.583	0.678
Rp(L)	0.576	0.674

Tabla 6.19. Recall y precisión normalizados para las distintas combinaciones de los modelos a largo plazo y corto plazo para la generación de resúmenes personalizados.

Los porcentajes de mejora sobre los valores medios de recall y precisión normalizados se resumen en la Tabla 6.20. La combinación de largo y corto plazo mejora la precisión respecto al largo plazo en un 1.6% y la mejora respecto al corto plazo en un 2.8%. Ambos resultados son estadísticamente significativos. El corto plazo da mejores resultados en precisión (1.2%) que el largo plazo, pero no de manera significativa.

Las mejoras en recall mantienen la tendencia aunque los porcentajes son menores. Se observa una pequeña diferencia significativa a favor de la combinación de largo y corto plazo respecto a la utilización del corto plazo en solitario (0.9%). También mejora significativamente la combinación frente al largo plazo en solitario (1.5%). Por último, la diferencia entre corto y largo (0.6%) no es significativa.

	% nP	% nR
Rp(LO) > Rp(L)	1.6	0.9
Rp(LO) > Rp(O)	2.8	1.5
Rp(O) > Rp(L)	1.2	0.6

Tabla 6.20. Porcentajes de mejora de las distintas combinaciones respecto a los valores medios de recall y precisión normalizados.

Los resultados obtenidos nos permiten retener la cuarta hipótesis (H4): los resúmenes, obtenidos usando sólo la heurística de personalización, son mejores si se utiliza una combinación de los modelos a largo y corto plazo, que si alguno de ellos se utiliza en solitario. Como resultado adicional, se observa que el modelo a corto plazo en solitario es mejor que el modelo a largo plazo en términos de precisión normalizada.

De aquí en adelante, cuando nos refiramos a resúmenes personalizados (Rp), estaremos refiriéndonos a la personalización obtenida mediante la combinación de los modelos a largo y corto plazo.

6.4.3.3. Experimento 5. Combinación de heurísticas para la generación de resúmenes.

Varias colecciones de evaluación diferentes se han generado para cada usuario, cada una de ellas consiste en el conjunto de resúmenes de las noticias originales obtenidos mediante la aplicación de cada uno de los métodos de generación de resúmenes indicados.

Las colecciones de resúmenes base (Rb) y de resúmenes genéricos (Rg) serán las mismas para todos los usuarios, puesto que no dependen de ningún aspecto del modelo de usuario, es decir, no están personalizadas. Esto permitirá evaluar a todos los usuarios a la vez.

Sin embargo, habrá una colección diferente por usuario de resúmenes personalizados (Rp) y de resúmenes genérico-personalizados (Rgp), puesto que éstas sí dependen del modelo del usuario. Habrá que evaluar por separado a cada usuario con respecto a su colección de resúmenes personalizados, por un lado, y respecto a su colección de resúmenes genérico-personalizados, por otro. Se han generado, aproximadamente 58 resúmenes distintos para una misma noticia.

En cada caso, se han calculado los valores de recall y precisión normalizados. Estos experimentos se han repetido durante los 14 días de evaluación. En total, se han generado aproximadamente 54950 resúmenes diferentes (58 resúmenes por noticia en media, 1099 noticias).

6.4.3.4. Resultados

Los resultados mostrados en la Tabla 6.21 muestran que los resúmenes personalizados ofrecen mejores resultados con respecto a la precisión normalizada de la información seleccionada que los resúmenes genérico-personalizados, que los resúmenes genéricos y que los resúmenes base. Los resúmenes genérico-personalizados son mejores que los resúmenes genéricos, y los resúmenes genéricos son mejores que los resúmenes base. También se puede observar en la misma tabla que los resúmenes personalizados son peores que las noticias completas (N) bajo el mismo criterio. En términos de recall normalizado, las tendencias se mantienen.

	N	Rp	Rgp	Rb	Rg
nP	0.603	0.593	0.584	0.581	0.577
nR	0.694	0.686	0.680	0.678	0.675

Tabla 6.21. Recall y precisión normalizados para los distintos tipos de resúmenes.

Los porcentajes de mejora sobre los valores medios de precisión y recall normalizados se resumen en la Tabla 6.22. Los resúmenes personalizados mejoran la precisión respecto a los resúmenes genérico-personalizados en un 1.5%, aunque no significativamente. Sin embargo, la mejora si es estadísticamente significativa frente a resúmenes base (2.0%) y resúmenes genéricos (2.8%). Las noticias completas dan mejor resultado que los resúmenes personalizados (1.7%), con una diferencia significativa. Los resúmenes genéricos-personalizados mejoran los resúmenes base en un ligero 0.5% no significativo y los resúmenes base ofrecen mejores resultados que los resúmenes genéricos en un 0.7% no significativo.

	N > Rp	Rp > Rgp	Rp > Pf	Rp > Rg	Rgp > Rb	Rb > Rg
% nP	1.7	1.5	2.0	2.8	0.5	0.7
% nR	1.0	1.0	1.2	1.6	0.2	0.4

Tabla 6.22. Porcentajes de mejora de los distintos tipos de resúmenes respecto a los valores medios de recall y precisión normalizados.

Las mejoras en recall mantienen la tendencia aunque los porcentajes son menores. Los resúmenes personalizados mejoran el recall respecto a los resúmenes genérico-personalizados en un 1.0% y también respecto a los resúmenes base en un 1.2%, aunque no significativa-

mente. Sin embargo, la mejora si es estadísticamente significativa frente a resúmenes genéricos (1.6%). Las noticias completas dan mejor resultado que los resúmenes personalizados (1.0%), con una diferencia significativa. Los resúmenes genéricos-personalizados mejoran los resúmenes base en un ligero 0.2% no significativo y los resúmenes base ofrecen mejores resultados que los resúmenes genéricos en un 0.4% no significativo.

Estos resultados nos permiten confirmar la hipótesis H5: Los resúmenes obtenidos usando sólo la heurística de personalización son mejores, con respecto a la información seleccionada por los usuarios, que los resúmenes obtenidos extrayendo las primeras frases del texto del documento completo.

También nos permiten retener, aunque en términos de recall normalizado los resultados no sean significativos, la hipótesis H6: Los resúmenes obtenidos utilizando únicamente la heurística de personalización son mejores con respecto a la información seleccionada por los usuarios, que los resúmenes obtenidos usando las heurísticas de generación de resúmenes genéricos y que los resúmenes obtenidos usando una combinación de heurísticas.

Por último, también queda confirmada parcialmente la hipótesis H7: Los resúmenes obtenidos usando sólo la heurística de personalización son peores que las noticias completas con respecto a la información seleccionada por los usuarios.

6.4.4. Evaluación cualitativa

Para contrastar los resultados de las evaluaciones cuantitativas realizadas sobre los procesos de selección, adaptación y presentación de resultados se realizó una evaluación cualitativa mediante el empleo de cuestionarios de evaluación que rellenaron los distintos usuarios del sistema. Las preguntas de estos formularios sirven para ver la opinión de los usuarios sobre los distintas partes del sistema y constituyen otro alternativa para medir la calidad del mismo.

Se realizaron dos cuestionarios sobre el sistema, una evaluación inicial para captar las opiniones de los usuarios antes de usar el sistema y una evaluación final para contrastar sus opiniones después de haber usado el sistema durante varios días. Estos cuestionarios se muestran en el Apéndice I.

La evaluación inicial debía ser rellenada por los usuarios justo después de registrarse en el sistema y definir su perfil de usuario. Estas operaciones se podían realizar conectándose a la página web donde estaba situado el sistema y navegando por las distintas páginas de registro y definición de perfil de usuario.

La evaluación final debía ser rellenada cuando un usuario dejara de usar del sistema, bien porque terminara su funcionamiento tras el 14º día, bien porque decidiera darse de baja en el sistema.

De los 106 usuarios del sistema, 90 rellenaron el cuestionario de evaluación inicial, mientras que sólo 38 rellenaron la evaluación final (Tabla 6.23). En total, hubo 35 usuarios que rellenaron tanto la evaluación inicial como la final.

De los 90 usuarios que realizaron la evaluación inicial, un 74.4% fueron alumnos, frente a un 18.9% de profesores, el 6.7% fueron otros profesionales de distintos tipos. Dentro de los profesores, el grupo mayoritario fue el de profesores que imparten clase en ingeniería informática (14.4% del total), seguido por profesores de otras carreras o profesores de instituto (3.3% del total). También participó un profesor de periodismo. Dentro de los alumnos el grupo más numeroso fue el de periodismo (38.9% del total), seguido por alumnos de la licenciatura de publicidad y relaciones públicas (26.7% del total) y por último, alumnos de

informática (8.9% del total). Del grupo de otros profesionales, 2 tenían relación con la informática y los otros 4 no tenían relación ni con la informática ni con el periodismo.

Otras estadísticas obtenidas en la evaluación inicial fueron que el número de mujeres fue 44 frente a 46 hombres y que el navegador utilizado por el 88.9% de los usuarios fue Internet Explorer, un 6.7% usó Mozilla, un 2.2% Netscape y un 2.2% otro diferente.

	Evaluación inicial	Evaluación final
Profesores informática	14.4	34.2
Otros profesores	4.4	2.6
Alumnos informática	8.9	10.5
Alumnos periodismo	38.9	26.3
Alumnos publicidad	26.7	18.4
Otros profesionales	6.7	7.9
Total usuarios	90	38

Tabla 6.23. Porcentajes de tipos de usuarios que realizaron la evaluación inicial y la evaluación final.

De los 38 usuarios que realizaron la evaluación final, un 55.2% fueron alumnos, frente a un 36.8% de profesores, el 7.9% fueron otros profesionales de distintos tipos. Dentro de los profesores, el grupo mayoritario fue el de profesores que imparten clase en ingeniería informática (34.2% del total), seguido por profesores de otras carreras o profesores de instituto (2.6% del total). Dentro de los alumnos el grupo más numeroso fue el de periodismo (26.3% del total), seguido por alumnos de la licenciatura de publicidad y relaciones públicas (18.4% del total) y por último, alumnos de informática (10.5% del total). Del grupo de otros profesionales, 1 tenía relación con la informática y los otros 2 no tenían relación ni con la informática ni con el periodismo.

Otra estadística obtenida fue que el número de hombres (65.8%) fue casi el doble que el de mujeres (34.2%).

En conclusión, los grupos son suficientemente heterogéneos y numerosos como para incluir distintos tipos de comportamiento frente al sistema que permitan extraer conclusiones significativas de sus evaluaciones.

6.4.4.1. Evaluación de la interfaz

En la evaluación inicial (Tabla 6.24) las opiniones de los usuarios sobre los componentes gráficos (pregunta 1) indican un grado de satisfacción mayoritariamente alto (50.6%), para un 4.5% fue muy alto, para un 34% fue regular, el 10.1% restante optó por bajo. Sin embargo a la pregunta de si el sistema era atractivo (pregunta 2), el 46.7% opinó que regular, frente a 38.9% de alto, 5.6% de muy alto, y 8.9% de bajo. En cuanto a la usabilidad (pregunta 3), el sistema fue considerado mayoritariamente como de fácil manejo (alto=60%), con 30% de muy alto y 10% de regular. También se consideró amigable (pregunta 4) para el usuario (alto=58.4%), con 7.9% de muy alto, frente a 30.3% de regular, y 3.4% de bajo. En cuanto a la gestión de contenidos (pregunta 5), el sistema de enlaces se consideró bueno en un 56.8% de las ocasiones, en un 8% fue muy bueno, frente a un 35.2% de regular. Por último, la ayuda proporcionada (pregunta 6) fue considerada como buena por el 67.4% de los usuarios, muy buena por el 11.6%, frente a un 18.6% de regular, y un 2.3% de mala.

	Evaluación inicial					Evaluación final				
Pregunta	muy alto	alto	regular	bajo	muy bajo	muy alto	alto	regular	bajo	muy bajo
1	4.5	50.6	34.8	10.1	0	2.6	55.3	42.1	0	0
2	5.6	38.9	46.7	8.9	0	5.3	39.5	47.4	7.9	0
3	30.0	60.0	10.0	0	0	26.3	60.5	13.2	0	0
4	7.9	58.4	30.3	3.4	0	2.6	60.5	34.2	2.6	0
5	8.0	56.8	35.2	0	0	2.6	71.1	21.1	5.3	0
6	11.6	67.4	18.6	2.3	0	7.9	47.4	42.1	2.6	0

Tabla 6.24. Porcentajes de usuarios sobre cada respuesta a cada pregunta sobre la evaluación de la interfaz, en la evaluación inicial y en la evaluación final.

Como se puede observar a vista de los resultados la impresión de los usuarios sobre la interfaz es bastante positiva, aunque para una minoría algunos aspectos tuvieron un grado de satisfacción bajo, ninguno de ellos lo consideró como muy bajo.

En la evaluación final (Tabla 6.24) se repitieron las cuestiones sobre el grado de satisfacción de la interfaz, teniendo en cuenta ahora los mensajes enviados a los usuarios con sus noticias personalizadas, además de la interfaz de registro y definición del perfil de usuario. Las opiniones sobre los componentes gráficos fueron de satisfacción alta (55.3%), 2.6% de muy alta, frente a 42.1% de regular. En cuanto a si el sistema era atractivo, el 47.4% opinaron regular, frente a 39.5% de alto, 5.3% de muy alto y 7.9% de bajo. Sobre la usabilidad, el sistema fue considerado fácil de usar en un grado alto por el 60.5%, con un 26.3% de muy alto y 13.2% de regular. El grado de amigabilidad también fue alto en un 60.5% de los casos, con un 2.6% de muy alto, frente a 34.2% de regular, y 2.6% de bajo. La gestión de contenidos fue considerada mayoritariamente en grado alto (71.1%), con un 2.6% de muy alto, frente a 21.1% de regular, y un 5.3% de bajo. Por último la ayuda se consideró buena en un 47.4% de los casos, muy buena en 7.9%, frente a 42.1% de regular, y 2.6% de mala.

En todos los casos se repitieron tendencias similares a la evaluación inicial, con un ligero incremento en las valoraciones (suma de alto y muy alto) sobre los componentes gráficos, grado de atracción del sistema y gestión de contenidos, del 54.9% al 57.9% en la primera, del 44.5% al 44.8% en la segunda. Sin embargo, se produjo un ligero decremento, del 90% al 86.8%, en la facilidad de uso, del 66.3% al 63.1%, en amigabilidad. Los cambios fueron más resaltables en la gestión de contenidos, incremento del 64.8% al 73.7%, y en el sistema de ayuda, decremento del 79.2% al 55.3%.

6.4.4.2. Valoración sobre las secciones, las categorías y las palabras clave

En la evaluación inicial las opiniones de los usuarios (Tabla 6.25) sobre la adecuación de las secciones para reflejar sus necesidades de información (pregunta 7) indican un grado de satisfacción mayoritariamente alto (66.7% de los usuarios), con un 11.5% de muy alto, para un 19.5% fue regular, frente a un 1.1% de bajo y otro 1.1% de muy bajo. En cuanto a las categorías (pregunta 8), la satisfacción fue alta en un 71.1% de los casos, con 3.3% de muy alta, 23.3% de regular, frente a un 1.1% de bajo y otro 1.1% de muy bajo. Las palabras clave (pre-

gunta 9) fueron consideradas satisfactorias en un grado alto para un 43.3% de los usuarios, con un remarcable 34.4% de muy alto, un 18.9% de regular, frente a un 3.3% de bajo.

	Evaluación inicial					Evaluación final				
Pregunta	muy alto	alto	regular	bajo	muy bajo	muy alto	alto	regular	bajo	muy bajo
7	11.5	66.7	19.5	1.1	1.1	5.3	52.6	34.2	7.9	0
8	3.3	71.1	23.3	1.1	1.1	0	62.2	27.0	8.1	2.7
9	34.4	43.3	18.9	3.3	0	24.2	30.3	27.3	9.1	9.1

Tabla 6.25. Porcentajes de usuarios sobre cada respuesta a cada pregunta sobre la valoración de secciones, generales y palabras clave, en la evaluación inicial y en la evaluación final.

Los usuarios consideran inicialmente que las palabras clave reflejan mejor sus necesidades de información que las secciones y las categorías, considerando estas dos últimas en un orden similar. En todo caso, la mayoría de los usuarios consideran que los 3 sistemas reflejan adecuadamente sus necesidades de información.

En la evaluación final (Tabla 6.25), el 52.6% de los usuarios consideran que las secciones han sido adecuadas para reflejar sus necesidades de información en un grado alto, con un 5.3% de muy alto, un 52.6% de alto, un 34.2% de regular, frente a un 7.9% de bajo. En cuanto a las categorías, la satisfacción fue alta en el 62.2% de los casos, con 27% de regular, frente a 8.1% de bajo y 2.7% de muy bajo. Las palabras clave fueron consideradas satisfactorias en un grado alto para un 30.3% de los usuarios, con un remarcable 24.2% de muy alto, un 27.3% de regular, frente a un 9.1% de bajo y otro 9.1% de muy bajo.

Considerando estos porcentajes, los usuarios consideraron después de usar el sistema que las categorías eran las que mejor reflejaban sus necesidades de información, seguidas muy de cerca por las secciones y por último, las palabras clave. Sin embargo, hay que matizar las opiniones sobre las palabras clave, mientras que el porcentaje de muy alto sigue siendo elevado (24.2%), por otro lado, el porcentaje de bajo y muy bajo es el de mayor valor de los 3 sistemas (18.2%). Esto es debido principalmente a que las palabras elegidas por los usuarios son, en general, pocas y muy específicas, haciendo que sea difícil que éstas aparezcan en las noticias para muchos usuarios, sin embargo, para aquellos para los que aparecen, dichas palabras son adecuadas para reflejar sus necesidades de información. Otro posible problema adicional con las palabras es el asociado a la polisemia, aunque en este caso afecta mucho menos debido a la especificidad de las palabras elegidas por los usuarios.

En todo caso, la mayoría de los usuarios consideran que los 3 sistemas reflejan adecuadamente sus necesidades de información, aunque los porcentajes son menores que en la evaluación inicial.

Uno de los factores que puede explicar la disminución en las opiniones de los usuarios en la evaluación final respecto a la evaluación inicial es la forma de utilización del sistema. Cuando los usuarios juzgaron el sistema éste no tenía el comportamiento normal en cuanto a que no le enviaba al usuario el número de noticias que él deseaba, sino todas las noticias. Esto es así porque había que recoger los juicios de relevancia sobre todas las noticias para poder evaluar posteriormente el sistema bajo distintas configuraciones. Bajo estas circunstancias no es tan sencillo darse cuenta de hasta qué punto el sistema personaliza bien la información porque hay que fijarse en el ranking en el que aparecen todas las noticias en lugar de evaluar simplemente si las 10 noticias que se reciben son o no interesantes.

Otra de las cuestiones (Tabla 6.26) que se formulaba se refería a la necesidad de incluir nuevas secciones a las ya existentes (pregunta 10). En la evaluación inicial, un 51.4% de los usuarios consideraba que se deberían introducir algunas secciones más para definir sus necesidades de información, con un 2.7% que opinaron muchas, un 40.5% pocas y un 5.4% ninguna. En la evaluación final, el 52.6% indicó ninguna, un 34.2% pocas, un 10.5% algunas y un 2.6% muchas. Cuando se les preguntó sobre cuáles serían esas nuevas secciones las respuestas fueron bastantes variadas: trabajo y empleo, tecnología, opinión, el tiempo, religión, televisión, ciencia, comunicación, información local y regional. En todo caso, se puede observar que los usuarios disminuyeron su necesidad de nuevas secciones después de haber usado el sistema.

	Evaluación inicial				Evaluación final			
Pregunta	muchas	algunas	pocas	ninguna	muchas	algunas	pocas	ninguna
10	2.7	51.4	40.5	5.4	2.6	10.5	34.2	52.6
11	28.9	60.5	0	10.5	5.3	15.8	32.4	55.9

Tabla 6.26. Porcentajes de usuarios sobre cada respuesta a cada pregunta sobre la introducción de nuevas secciones y categorías, en la evaluación inicial y en la evaluación final.

En la evaluación inicial, un 60.5% de los usuarios consideraba que se deberían introducir algunas categorías más para definir sus necesidades de información (pregunta 11), con un 28.9% que opinaron muchas, un 10.5% ninguna. En la evaluación final, el 55.9% indicó ninguna, un 32.4% pocas, un 15.8% algunas y un 5.3% muchas. Cuando se les preguntó sobre cuáles serían esas nuevas categorías las respuestas fueron bastantes variadas: curiosidades, humor, fútbol femenino, publicidad, espectáculos, religión, sucesos, conflictos bélicos, famosos, bolsa; dos usuarios echaron en falta categorías más específicas. En todo caso, se puede observar que los usuarios disminuyeron su necesidad de nuevas categorías después de haber usado el sistema.

En cuanto a cambios en las secciones y las categorías a lo largo de la utilización del servicio, sólo hubo 3 usuarios (7.9%) que cambiaron 1 vez de secciones, 1 (2.8%) que cambió de categorías y ninguno cambió de palabras clave. La gran mayoría de los usuarios no necesitaron cambiar sus perfiles a largo plazo, aunque algunos de ellos manifestaron en las preguntas abiertas el desconocimiento de esa posibilidad.

También se preguntó en la evaluación final sobre la selección (Tabla 6.27), por parte del sistema, de documentos que pertenecieran a las secciones (pregunta 12), categorías (pregunta 13) o palabras clave (pregunta 14) escogidas por el usuario antes que documentos que no pertenecieran a ellas. Para las secciones se obtuvo un grado alto en un 54.1% de los casos, con un 8.1% de muy alto, un 29.7% de regular, frente a un 8.1% de bajo. Para las categorías, hubo un 45.5% de alto, un 6.1% de muy alto, un 36.4% de regular, frente a un 9.1% de bajo y un 3% de muy bajo. Para las palabras clave, se obtuvo un 33.3% de alto, un 12.1% de muy alto, un 24.2% de regular, frente a un 24.2% de bajo y un 6.1% de muy bajo.

Pregunta	muy alto	alto	regular	bajo	muy bajo
12	8.1	54.1	29.7	8.1	0
13	6.1	45.5	36.4	9.1	3.0
14	12.1	33.3	24.2	24.2	6.1

Tabla 6.27. Porcentajes de usuarios sobre cada respuesta a cada pregunta sobre la selección de documentos según los distintos sistemas de referencia, en la evaluación final.

La mayoría de los usuarios consideran que las secciones son las que permiten seleccionar más adecuadamente los documentos. Resultados un poco inferiores aparecen para las categorías, pero siendo todavía mayoritarias las valoraciones positivas. Sin embargo, se muestran más escépticos con respecto a las palabras clave (30.3% de bajo y muy bajo). Esto, de nuevo, se puede deber tanto al nivel de especificidad de las palabras clave como a los posibles problemas de polisemia que puedan aparecer en algunas de ellas.

En la evaluación final se plantearon además otras cuestiones sobre las palabras clave (Tabla 6.28). En primer lugar, se preguntó sobre el grado de correspondencia entre los documentos recuperados según las palabras clave y las necesidades de información del usuario (pregunta 15), obteniéndose un 40.6% de regular, un 25% de alto, un 15.6% de muy alto, frente a un 12.5% de bajo y un 6.3% de muy bajo. Otra pregunta fue sobre la claridad de las razones de recuperación de los documentos (pregunta 16), en este caso, el grado de claridad fue regular en el 43.8% de los casos, con un 31.3% de alto, un 9.4% de muy alto, frente a un 9.4% de bajo y un 6.3% de muy bajo. La tercera pregunta versó sobre la influencia de la especificidad de las palabras en la recuperación de los documentos (pregunta 17). El 41.9% optó por un grado de influencia alto, con un 3.2% de muy alto, un 38.7% de regular, frente a un 9.7% de bajo y un 6.5% de muy bajo. Por último, los usuarios consideraron útil la utilización de algún instrumento para la introducción de palabras relacionadas (pregunta 18) en un 34.3% de los casos, con un 14.3% de muy alto, un 20% de regular y un 32.4% de bajo.

Pregunta	muy alto	alto	regular	bajo	muy bajo
15	15.6	25.0	40.6	12.5	6.3
16	9.4	31.3	43.8	9.4	6.3
17	3.2	41.9	38.7	9.7	6.5
18	14.3	34.3	20.0	31.4	0

Tabla 6.28. Porcentajes de usuarios sobre cada respuesta a cada pregunta sobre las palabras clave, en la evaluación final.

Es claro que la utilización de las palabras clave en solitario para la definición del perfil de usuario habría llevado a un descontento mayoritario en el uso del sistema, aunque para alrededor de un 40% de los usuarios parece que la utilización de las palabras clave fue satisfactoria. En todo caso, aproximadamente un 50% de ellos reconoce la utilidad de algún instrumento que le mostrara palabras relacionadas para poder mejorar la definición de su perfil de usuario.

Finalmente se preguntó en la evaluación final sobre la preferencia de sistema para definir los intereses del usuario (Tabla 6.29 – pregunta 19): un 43.2% eligió las palabras clave, un 29.7% las categorías y un 24.3% las secciones. Hubo 1 usuario que eligió otras opciones, indicando la realimentación como su preferencia para definir sus intereses.

Pregunta	Secciones	Categorías	Palabras clave
19	24.3	29.7	43.2

Tabla 6.29. Porcentajes de usuarios sobre la preferencia en el sistema de referencia, en la evaluación final.

Curiosamente, a pesar de los problemas en la selección con las palabras clave, los usuarios sienten que es la mejor forma de definir sus intereses debido a la especificidad del sistema de referencia frente a los sistemas de secciones y categorías donde la definición de intereses es mucho más amplia.

6.4.4.3. Valoración sobre la medida de la relevancia de las noticias

Se mostraron al usuario los siguientes criterios para decidir sobre la medida de la relevancia de las noticias: criterio 1, la perspectiva y el planteamiento; criterio 2, la profundidad y la cantidad de información; criterio 3, el estilo; criterio 4, la novedad; criterio 5, la utilidad; criterio 6, la relación con su perfil de usuario; criterio 7, la relación con sus necesidades de información y con sus temas de interés; criterio 8, la capacidad para añadir nuevo conocimiento frente a otros documentos relacionados; criterio 9, la cercanía y grado de motivación desde un punto de vista emocional; criterio 10, la cercanía desde un punto de vista geográfico; criterio 11, la cercanía y familiaridad con el contenido expuesto; criterio 12, la cercanía y familiaridad con el lenguaje empleado.

En la evaluación inicial los criterios que fueron seleccionados para la decisión sobre la relevancia de las noticias (Tabla 6.30) por más de un 50% de los usuarios fueron, en orden: la relación con sus necesidades de información y con sus temas de interés (criterio 7, 87.8%), la relación con su perfil de usuario (criterio 6, 65.6%), la utilidad (criterio 5, 57.8%) y la novedad (criterio 4, 54.4%). El menos seleccionado fue el estilo (criterio 3, 17.8%), seguido de por la cercanía y grado de motivación desde un punto de vista emocional (criterio 9, 20%), la cercanía y familiaridad con el contenido expuesto (criterio 10, 26.7%) y la cercanía y familiaridad con el lenguaje empleado (criterio 12, 28.9%).

Criterio	Evaluación inicial	Evaluación final
1	41.1	44.7
2	47.8	15.8
3	17.8	15.8
4	54.4	57.9
5	57.8	55.3
6	65.6	68.4
7	87.8	84.2
8	46.7	57.9
9	20.0	28.9
10	26.7	31.6
11	38.9	28.9
12	28.9	18.4

Tabla 6.30. Porcentajes de usuarios sobre la medida de la relevancia de las noticias, en la evaluación inicial y en la evaluación final.

En la evaluación final los criterios que fueron seleccionados para la decisión sobre la relevancia de las noticias por más de un 50% de los usuarios fueron, en orden: la relación con sus necesidades de información y con sus temas de interés (criterio 7, 84.2%), la relación con su perfil de usuario (criterio 6, 68.4%), la novedad (criterio 8, 57.9%), la capacidad para añadir nuevo conocimiento frente a otros documentos relacionados (criterio 4, 57.9%) y la utilidad (criterio 5, 55.3%). Los menos seleccionados fueron el estilo (criterio 3, 15.8%) y la profundidad y la cantidad de información (criterio 2, 15.8%), seguido de por la cercanía y familiaridad con el lenguaje empleado (criterio 12, 18.4%), la cercanía y grado de motivación desde un punto de vista emocional (criterio 11, 28.9%), la cercanía y familiaridad con el contenido expuesto (criterio 9, 28.9%).

Los criterios son más o menos coincidentes excepto la capacidad para añadir nuevo conocimiento frente a otros documentos relacionados que incrementa mucho su relevancia en la evaluación final. En realidad, este criterio es el segundo que más varía pasando de un 46.7% en la evaluación inicial a un 57.9% en la final. El incremento de este criterio se debe a que en las noticias diarias de un periódico hay varias que tratan sobre lo mismo. Estas noticias tratan aspectos distintos del mismo tema pero repitiendo mucha información, por tanto el usuario considera como más relevante la noticia si le aporta nuevo conocimiento y no si le repite la información que ya conoce.

Por otro lado, el criterio que más varía es la profundidad y la cantidad de información, pasando de un 47.8% a un 15.8%, esto es debido a que el usuario lo que quiere es información concisa sobre las noticias, más que profundidad y cantidad de información. Este aspecto se ve agudizado en la evaluación del sistema al tener los usuarios que revisar todas las noticias de un mismo día.

Por otro lado, en la evaluación inicial, el nivel de interés por la información que iba a ser recibida (Tabla 6.31 – pregunta 20) era alto en el 60% de los casos, con un 25.6% de muy alto, un 13.3% de regular, frente a un 1.1% de bajo. Después de usar el sistema, ese nivel de interés pasa a ser alto en un 45.9% de las ocasiones, muy alto en un 8.1%, regular en un 37.8%, frente a un 8.1% de bajo.

	Evaluación inicial					Evaluación final				
Pregunta	muy alto	alto	regular	bajo	muy bajo	muy alto	alto	regular	bajo	muy bajo
20	25.6	60.0	13.3	1.1	0.0	8.1	45.9	37.8	8.1	0.0

Tabla 6.31. Porcentajes de usuarios sobre el interés de los usuarios por la información recibida, en la evaluación final.

Los usuarios se muestran menos interesados a lo largo del funcionamiento del sistema, esto puede ser debido a la cantidad de trabajo exigida en la evaluación diaria.

Por último, se le preguntó a los usuarios, en la evaluación final, por la información relacionada con la noticia que habían utilizado para decidir sobre la relevancia de cada noticia (Tabla 6.32). Se obtuvo que el 89.2% usaron el título muchas veces y un 10.8% algunas. La sección fue utilizada algunas veces por el 45.9%, muchas por el 29.7%, pocas por el 13.5% y ninguna por el 10.8%. La relevancia la usaron algunas veces el 35.1% de los usuarios, pocas el 24.3%, ninguna el 21.6% y muchas el 18.9%. El resumen lo utilizaron muchas veces el 48.6% de los usuarios, algunas veces el 29.7% y pocas el 21.6%. Por último, la noticia completa fue utilizada pocas veces por el 51.4% de los usuarios, algunas veces por el 29.7% y ninguna por el 18.9%.

Información	Muchas	Algunas	Pocas	Ninguna
Título	89.2	10.8	0.0	0.0
Sección	29.7	45.9	13.5	10.8
Relevancia	18.9	35.1	24.3	21.6
Resumen	48.6	29.7	21.6	0.0
Noticia completa	0.0	29.7	51.4	18.9

Tabla 6.32. Porcentajes de usuarios sobre la preferencia de los usuarios en la información relacionada con la noticia utilizada para determinar la relevancia, en la evaluación final.

Los usuarios utilizaron el título para decidir sobre la relevancia de las noticias mucho más que las otras informaciones relacionadas. En segundo lugar, lo más utilizado fue el resumen. De hecho estas dos informaciones fueron utilizadas siempre por todos los usuarios. A continuación se utilizaron la sección y la relevancia y por último la noticia completa.

Esta utilización del resumen con mucho más frecuencia que la noticia completa, e incluso más que la sección y la relevancia, justifica el esquema de presentación de resultados propuesto en este trabajo.

6.4.4.4. Valoración sobre los resúmenes

En la evaluación final se preguntaron diversas cuestiones sobre los resúmenes que se comentan a continuación (Tabla 6.33). Los usuarios indicaron que los resúmenes eran de calidad alta (pregunta 21) en un 72.2% de los casos, con un 11.1% de muy alto, un 11.1% de regular, frente a un 5.6% de muy bajo. En cuanto a la coherencia y claridad (pregunta 22), el 67.6% dio una valoración alta, un 13.5% la dio muy alta, un 13.5% regular, frente a un 2.7% que la dio baja y otro 2.7% muy baja. La capacidad del sistema para evitar redundancias (pregunta 23) fue valorada alta por un 61.1% de los usuarios, muy alta por un 8.3%, regular por un 27.8%, frente a un 2.8% de baja. La adaptación del resumen al perfil de usuario (pregunta 24)

fue considerada en un grado alto por un 59.5% de los usuarios, con un 5.4% de muy alto, un 27% de regular, frente a un 5.4% de bajo y un 2.7% de muy bajo. La adaptación de los resúmenes a las necesidades de información (pregunta 25) fue alta en un 62.2% de los casos, muy alta en un 8.1%, regular en un 18.9%, baja en un 5.4% y muy baja en otro 5.4%. Los resúmenes reflejan el contenido de los documentos (pregunta 26) en un grado alto para el 64.9% de los usuarios, en un grado muy alto para el 16.2%, regular para el 13.5%, bajo para el 2.7% y muy bajo para el 2.7%. Por último, el 89.5% de los usuarios consideran que los principales componentes de la noticia están representados en la noticia. El otro 10.5% indicó que, a veces, los resúmenes eran excesivamente breves para contenerlos.

Pregunta	muy alto	alto	regular	bajo	muy bajo
21	11.1	72.2	11.1	5.6	0.0
22	13.5	67.6	13.5	2.7	2.7
23	8.3	61.1	27.8	2.8	0.0
24	5.4	59.5	27.0	5.4	2.7
25	8.1	62.2	18.9	5.4	5.4
26	16.2	64.9	13.5	2.7	2.7

Tabla 6.33. Porcentajes de usuarios sobre cada respuesta a cada pregunta sobre los resúmenes, en la evaluación final.

Los usuarios consideran mayoritariamente que los resúmenes son de calidad, coherentes y claros, y que reflejan el contenido y los principales componentes de la noticia. También consideran mayoritariamente, aunque en un porcentaje menor, que los resúmenes no contienen redundancias, se adaptan al perfil y a las necesidades del usuario.

Esta evaluación positiva justifica el método de selección de frases para la construcción de resúmenes como válido para la presentación de resultados a pesar de los problemas iniciales comentados referentes a sus problemas de claridad, coherencia y redundancia. De hecho el porcentaje menor en la redundancia puede ser debido a esto.

Por otro lado, los porcentajes menores en la adaptación al perfil y a las necesidades de información pueden ser debidos a que el usuario examina muchas noticias que no están relacionadas ni con su perfil ni con sus necesidades, por lo tanto, el resumen no se puede adaptar a algo con lo que no tiene relación.

6.4.4.5. Valoración sobre la selección y la adaptación

A continuación vamos a presentar la estimación de los usuarios respecto de los procesos de selección y adaptación obtenida a partir del cuestionario de evaluación final (tabla 6.34).

Pregunta	muy alto	alto	regular	bajo	muy bajo
27	8.1	51.4	35.1	5.4	0
28	8.1	27.0	56.8	8.1	0
29	5.4	37.8	40.5	16.2	0
30	2.9	5.7	28.6	34.3	28.6

Tabla 6.34. Porcentajes de usuarios sobre valoraciones sobre la selección y la adaptación, en la evaluación final.

El 51.4% de los usuarios consideró que el sistema mostraba antes las noticias correspondientes a sus necesidades de información que otras noticias (pregunta 27) en un grado alto, el 8.1% lo consideró muy alto, el 35.1% regular, frente al 5.4% de bajo. La adaptación con el tiempo del sistema (pregunta 28) fue considerada como regular por el 56.8% de los usuarios, alta por el 27.0%, muy alta por el 8.1%, frente a baja por el 8.1%. La adaptación a los juicios realizados por los usuarios (pregunta 29) fue considerada como regular por el 40.5%, alta por el 37.8%, muy alta por el 5.4%, frente a baja por el 16.2%. Por último, el cambio en las necesidades de información a lo largo de la utilización del sistema (pregunta 30) fue bajo para el 34.3%, muy bajo para el 28.6%, regular para el 28.6%, alto para el 5.7%, y muy alto para el 2.9%.

Así pues, los usuarios consideran que la adaptación del sistema a sus juicios de relevancia es mayor que a sus necesidades de información, aunque esas adaptaciones son satisfactorias para menos del 50% de los usuarios. Por otro lado, la mayoría de los usuarios no cambian sus necesidades de información a lo largo de la utilización del sistema. Por último, fijándose en los resultados de la pregunta 14, se puede observar que los usuarios juzgan mayoritariamente que reciben las noticias que les interesan antes que las que no les interesan, con lo cual podemos concluir que los procesos de selección y adaptación se valoran mayoritariamente de manera positiva, puesto que el hecho de que se produzca esta valoración es debido a que dichos procesos funcionan adecuadamente.

6.4.4.6. Estimación global del sistema

Las valoraciones globales del sistema en la evaluación final muestran (tabla 6.35) que el nivel general de satisfacción y confianza en el sistema (pregunta 31) fue considerado alto por el 52.8% de los usuarios, muy alto por el 13.9% y regular por el 33.3%. Ningún usuario emitió la valoración de bajo o muy bajo. El sistema resolvió las necesidades de información (pregunta 32) en un grado alto al 63.3% de los usuarios, regular al 33.3% y bajo para un 3.3%. Los usuarios opinaron sobre la medida de personalización del sistema (pregunta 33) optando por una medida regular el 47.2%, alta el 44.4% y muy alta el 8.3%. Ninguno eligió baja o muy baja.

Pregunta	muy alto	alto	regular	Bajo	muy bajo
31	13.9	52.8	33.3	0	0
32	0	63.3	33.3	3.3	0
33	8.3	44.4	47.2	0	0

Tabla 6.35. Porcentajes de usuarios sobre las estimaciones globales del sistema, en la evaluación final.

En conclusión, la mayoría de los usuarios juzgan el sistema positivamente. Y de nuevo, entendemos que los porcentajes no son mayores debido a la tarea de evaluación de todas las noticias durante todos los días.

6.4.4.7. Preguntas abiertas

Tanto en la evaluación inicial como en la evaluación final se han realizado una serie de cuestiones generales sobre el sistema de las cuales se extrae a continuación un resumen de los comentarios más interesantes. Hay que recalcar que muchos usuarios no respondieron a estas preguntas abiertas, esto es, las dejaron en blanco.

Se preguntó sobre las características más importantes del sistema obteniendo los siguientes comentarios: en la evaluación inicial los comentarios más comunes hacían hincapié en la capacidad de recibir información adaptada al usuario (33.3%), su sencillez y facilidad de manejo (17.7%). Algunos usuarios resaltaron la utilización de palabras clave junto a secciones y categorías para definir el perfil de usuario (6.3%), la capacidad de aprendizaje (3.1%), que sea eficaz (3.1%), la innovación (3.1%), la gratuidad y flexibilidad de contenidos (1%). En la evaluación final la mayoría de los comentarios fueron similares, capacidad de recibir información personalizada según los intereses definidos por el usuario (39.5%), facilidad de uso (13.2%), personalización con palabras clave (7.9%), pero también se hicieron varios comentarios sobre la adaptación a los intereses del usuario a lo largo del tiempo (7.9%) y la utilización de los resúmenes (5.3%). Hubo opiniones sueltas que optaron por las instrucciones a seguir (2.6%) y la eficacia (2.6%).

En segundo lugar se preguntó a los usuarios sobre los elementos que echaba en falta en el sistema. En la evaluación inicial la respuesta más común fue ninguno (12.5%), aunque también hubo varios usuarios que echaban de menos un diseño de la interfaz más atractivo (8.3%); algunos usuarios solicitaban noticias de otros periódicos (5.2%), más ayuda a la hora de rellenar el perfil de usuario (4.2%), alguna sección o categoría más (2.1%), secciones o categorías más específicas (2.1%), opiniones sueltas echaban en falta más publicidad (1%), un buscador interno (1%), uso de expresiones en lugar de palabras (1%) y utilización real (1%), es decir, pocas noticias. En la evaluación final, la opinión que más se emitió era respecto al número de noticias enviadas en el mensaje (23.7%), los usuarios solicitaban recibir menos noticias, sólo las que realmente les interesaban, evidentemente ese sería el funcionamiento real del sistema, pero el propósito del experimento era además obtener una colección de juicios sobre todas las noticias, por lo tanto no se les podía enviar menos. Algunos usuarios echaban en falta noticias de otros periódicos (10.5%), diseño de la interfaz más atractivo (7.9%), fotos (7.9%), más información sobre el usuario en los mensajes (5.3%), alternativas de clasificación de las noticias (no sólo relevancia) (5.3%). Por otro lado, hubo opiniones que pedían que la selección también se pudiera hacer por el autor de la noticia (2.6%), que se mostrara la categoría o categorías a las que pertenecía cada noticia (2.6%) y que se pudieran agrupar las noticias relacionadas (2.6%).

En tercer lugar se preguntó a los usuarios sobre las necesidades de información que tenían antes y después de usar el sistema. Lo primero se preguntó en la evaluación inicial y lo segundo en la evaluación final. En general, las necesidades de información se mantuvieron constantes antes y después de usar el sistema, aunque en algunos se especificaran algo más estas necesidades después de usar el sistema. En cuanto al tipo de necesidades indicadas por los usuarios, casi todas trataban sobre los temas indicados en sus perfiles de usuario, algunas otras indicaban interés por los resúmenes sobre noticias de actualidad en general.

En cuarto lugar, los usuarios indicaron el tipo de información que les interesa más desde un punto de vista del tipo de documento (reportaje, crónica, editorial, etc.), antes y después de usar el sistema. La mayoría de los intereses indicaron su preferencia por los reportajes, seguido por las crónicas y los artículos de opinión. Algunos especificaban un poco más, indicando su preferencia por crónicas de actualidad y reportajes sobre, en algunos casos, investigación o noticias científico-técnicas, y en otros casos, sobre las noticias más relacionadas con las categorías y palabras clave seleccionadas en el perfil de usuario. No hubo diferencia significativa entre las respuestas emitidas en la evaluación final y las emitidas en la evaluación inicial.

Además se preguntó en la evaluación final a los usuarios sobre su creencia sobre si el sistema permitía o no la interacción con el usuario. La mayoría de ellos (más del 85% de los que

contestaron la pregunta) respondieron afirmativamente, indicando que la realimentación sobre las noticias permitía la interacción entre el usuario y el sistema. Sin embargo, algunos de los usuarios se mostraban críticos porque el sistema siempre les mostraba una serie de noticias que no le interesaban nada, esto es inevitable porque el funcionamiento del sistema estaba pensado para mandar todas las noticias a los usuarios para que pudieran emitir su opinión sobre cada una de ellas. Algunos otros echaban de menos poder ver el efecto de la realimentación sobre su perfil de usuario para poder entender mejor al sistema y en cualquier caso, corregirlo si no definiera correctamente sus intereses.

Por último se recogieron comentarios generales de los usuarios en un apartado tanto en la evaluación inicial como en la final. Los escasos comentarios de la evaluación inicial trataron principalmente sobre la expectación antes del inicio del funcionamiento del sistema. También hubo un par de comentarios sobre la mejora de la interfaz gráfica. En la evaluación final se recogieron comentarios positivos del funcionamiento del sistema aunque de nuevo varios usuarios volvieron a hacer hincapié sobre el problema de recibir muchas noticias, los usuarios se sentían un poco saturados con tanta información, y sobre la mejora del interfaz de usuario para que fuera más vistoso. Un comentario interesante sugería la posibilidad de resumir todas las noticias que trataran sobre un mismo tema en una único resumen para evitar leer la misma o similar información varias veces. Esto ocurría con, al menos, un par de grupos de 3 a 5 noticias, por día. Desde luego hacer un resumen multidocumento sería una mejora indiscutible del sistema. Otro comentario interesante trataba sobre el resumen personalizado de noticias poco relevantes para el usuario según el criterio de selección. El hecho de construir un resumen con las frases más relevantes al perfil de usuario de estas noticias puede dar lugar a un resumen que puede crear confusión al usuario sobre su interés sobre la noticia. En todo caso, este problema no se presentaría en un funcionamiento real del sistema en el que se minimizaría el número de noticias poco relevantes que recibiría el usuario.

6.4.5. Conclusiones del sistema de personalización 2.0

El sistema de personalización 2.0 se ha evaluado respecto a los 3 procesos de personalización mediante la utilización de la colección 2.0. El modelo de usuario contiene secciones, categorías y palabras clave.

El número de usuarios implicado en la evaluación del sistema y su distribución en grupos heterogéneos permite extraer conclusiones mucho más significativas que las obtenidas para los sistemas de personalización anteriores.

Uno de los objetivos de esta evaluación consiste en la comparación entre los resultados de la evaluación cualitativa y la evaluación cuantitativa con respecto a los distintos procesos de personalización.

En términos generales, los usuarios destacan la conveniencia de la combinación de varios sistemas para definir el perfil del usuario, aunque curiosamente, en un nivel inicial, los usuarios recalcan el valor de las palabras clave, hecho que, con el uso del sistema se modifica en gran medida.

En lo que afecta a la selección y a la adaptación de contenidos, se puede señalar que la evaluación de los usuarios coincide con los resultados de la evaluación cuantitativa. En primer lugar, existe coincidencia en que la utilización de todos los sistemas de referencia es bastante satisfactoria. En la evaluación cualitativa las valoraciones son claramente positivas sobre todos los métodos de clasificación y en la evaluación cuantitativa los resultados prove-

nientes de combinaciones son significativamente mejores, tanto en selección como en adaptación, que los que utilizan métodos en solitario.

Por otro lado, en cuanto a la utilización de cada método de clasificación en particular los resultados requieren una discusión más detallada. En la evaluación cualitativa, las opiniones sobre la adecuación de cada método varían desde las valoraciones iniciales donde la preferencia es por las palabras clave, seguidas por secciones y categorías, estas últimas con valoraciones similares, a las valoraciones finales donde las preferencias son secciones, palabras clave y categorías, aunque con puntuaciones similares. Por otro lado, a la hora de mostrar documentos correspondientes a los intereses reflejados en cada sistema de referencia antes que documentos que no correspondan, los usuarios valoran mejor las secciones, seguidas de categorías y por último, palabras clave. Y a la hora de indicar qué sistema de referencia prefieren después de usar el sistema, el primer lugar es para las palabras clave, seguido de categorías y secciones. Sin embargo, en la evaluación cuantitativa los peores resultados proceden claramente de las palabras clave, seguidas por categorías y secciones, siendo las categorías mejores en precisión y las secciones mejores en recall.

Como se puede observar, las opiniones sobre las palabras clave tienen un comportamiento peculiar. Los usuarios, aunque modifiquen a la baja su visión sobre la idoneidad de las palabras clave para reflejar sus intereses, siguen prefiriendo este sistema, lo que puede reflejar la preferencia por determinar por sí mismos sus intereses de manera lo más específica posible.

En cuanto a la medida de la relevancia de las noticias, se ha encontrado especial preferencia por criterios de interés, es decir, que la noticia fuera relevante para el usuario, seguido por criterios de novedad. Por otro lado, el estilo, la profundidad y la cercanía han sido los criterios menos utilizados por los usuarios. El criterio asociado a añadir nuevo conocimiento se ha incrementado en aproximadamente en un 10% debido a la presencia en las noticias de varios artículos que tratan sobre el mismo tema. Por otro lado, la profundidad y cantidad de información ha disminuido en un 32% debido al interés del usuario en información lo más concisa posible.

Los usuarios utilizaron siempre el título y el resumen para decidir sobre la relevancia de las noticias. Esta utilización del resumen con mucho más frecuencia que la noticia completa, e incluso más que la sección y la relevancia, justifica el esquema de presentación de resultados propuesto en este trabajo.

En cuanto a valoraciones sobre resúmenes, los usuarios consideraron mayoritariamente los resúmenes de calidad, coherentes, claros y que reflejan el contenido y los principales componentes de la noticia. En un porcentaje menor, hubo valoraciones positivas sobre el contenido de redundancias, adaptación al perfil y las necesidades del usuario. Estas valoraciones justifican la utilización del método de selección de frases para la construcción de resúmenes

En conclusión, la mayoría de los usuarios juzgan de manera positiva tanto los procesos de personalización (selección, adaptación y presentación) como el sistema en su globalidad. Sin embargo, los porcentajes no son mayores debido a la tarea de evaluación de todas las noticias durante todos los días.

Por último, resaltar algunas mejoras sugeridas por los usuarios en los comentarios, como mejorar la interfaz de usuario, envío de menos noticias, utilización de varios periódicos como fuentes de información o generación de resúmenes multidocumento.

6.5. Resumen y conclusiones del capítulo

En este capítulo se han presentado los distintos sistemas de personalización de noticias desarrollados en la tesis, así como los datos obtenidos de su evaluación utilizando las colecciones de evaluación del capítulo anterior.

En los experimentos preliminares se ha podido observar que las contribuciones realizadas por los distintos métodos de clasificación muestran la complejidad de los factores que intervienen en el sistema. También se ha podido observar los resultados altos de recall y precisión obtenidos con las secciones, los resultados bajos obtenidos por las categorías y los resultados bajos en recall y altos en precisión obtenidos para las palabras clave.

El problema en la normalización de las relevancias obtenidas a partir de cada sistema de clasificación hace que la relevancia proveniente de las distintas contribuciones no sea comparable directamente, permitiendo que unos sistemas de referencia tengan mucho más peso que otros. En particular la influencia de las secciones es muy superior a la de categorías y palabras clave.

La representación de las categorías se ha obtenido sólo con la información de las páginas de primer nivel de Yahoo! España. Esto ha llevado a una representación pobre para algunas categorías y por tanto a que la relevancia proveniente de este método de clasificación sea baja. El problema se agudiza debido a que los resultados no se normalizan cuando se combinan los distintos métodos de clasificación.

La evaluación de los experimentos preliminares estaba limitada por que los juicios de relevancia eran generados por expertos humanos. Estos juicios no tienen en cuenta las necesidades exactas de cada usuario, sino que simplemente se basan en la información contenida en los modelos de usuario. Sin embargo, los sistemas de personalización 1.0 y 2.0 si se evalúan con respecto a juicios de relevancia generados por los usuarios, lo cual permite una mejor evaluación al utilizar juicios reales asociados a las necesidades de información de cada uno de los usuarios.

El sistema de personalización 1.0 se ha evaluado respecto a los 3 procesos de personalización mediante la utilización de la colección 1.0. El modelo de usuario sólo contiene secciones y palabras clave, y se realiza normalización de contribuciones cuando son combinadas.

La evaluación del proceso de selección de contenidos ha mostrado que la combinación de secciones y palabras clave ofrece mejores resultados que la utilización de secciones o palabras clave por separado. Por otro lado, se ha observado que las secciones ofrecen mejores resultados que las palabras clave.

De la evaluación del proceso de adaptación del modelo de usuario se ha obtenido que la combinación de largo y corto plazo es mejor que la utilización de los modelos en solitario. El modelo a largo plazo en solitario es mejor que el modelo a corto plazo en solitario. También se ha obtenido que los mejores resultados en la combinación de los modelos a largo y corto plazo se producen cuando se combinan secciones y palabras clave en el modelo a largo plazo, como cabría esperar de los resultados obtenidos en los experimentos sobre la selección de contenidos. La opción de sólo secciones es mejor que sólo palabras clave.

La evaluación de la presentación de resultados permite concluir que los resúmenes personalizados utilizando una combinación de los modelos a largo y corto plazo es mejor que los otros tipos de resúmenes. Por otro lado, las noticias completas ofrecen una ligera mejora no significativa frente a los resúmenes personalizados, lo cual quiere decir que la pérdida de información para el usuario es muy pequeña con este tipo de resúmenes.

El sistema de personalización 2.0 se ha evaluado respecto a los 3 procesos de personalización mediante la utilización de la colección 2.0. El modelo de usuario contiene secciones, categorías y palabras clave. En este proceso se ha efectuado tanto una evaluación cuantitativa como una cualitativa.

Los mejores resultados, tanto en selección como en adaptación, se obtienen con combinación de sistemas. Las opiniones de los usuarios coinciden al dar valoraciones altas a todos los mecanismos de selección. A la hora de valorar individualmente los sistemas de referencia, los usuarios terminan seleccionando las palabras clave como mejor método para definir sus intereses, a pesar de reconocer peores valores en la selección de documentos a partir del mismo. Esto refleja la preferencia de los usuarios por determinar sus intereses de manera lo más específica posible.

Los criterios de interés son los más seleccionados por los usuarios, mientras que la profundidad y la cercanía son los menos elegidos. Además el hecho de añadir nuevo conocimiento ha sido el criterio que más ha aumentado su valoración tras la utilización del sistema.

La utilización del resumen como criterio para decidir la relevancia de una noticia justifica el esquema de presentación de resultados propuesto en este trabajo. Además las valoraciones altas sobre las distintas características de los resúmenes justifican la utilización del método de selección de frases.

En conclusión, la mayoría de los usuarios juzgan de manera positiva tanto los procesos de personalización (selección, adaptación y presentación) como el sistema en su globalidad. Sin embargo, los porcentajes no son mayores debido a la tarea de evaluación de todas las noticias durante todos los días.

Capítulo 7
DISCUSIÓN DE RESULTADOS

7.1. Introducción

En este capítulo se va a discutir sobre las distintas propuestas presentadas a lo largo de la tesis, comparándolas entre sí y con el estado del arte. También se presentará una extrapolación del modelo de personalización a un ámbito multilingüe.

En el apartado 7.2 se realizará una comparación de los resultados obtenidos con las distintas versiones de sistemas de personalización presentadas en el capítulo anterior.

En el apartado 7.3 se efectuará una comparación de la propuesta general presentada en la tesis con las propuestas más similares que aparecen en el estado del arte.

En el apartado 7.4 se presentará una propuesta de personalización multilingüe como extesnión del modelo de personalización monolingüe presentado hasta el momento.

Por último, en el apartado 7.5 se mostrará un resumen y las conclusiones del capítulo.

7.2. Comparación de resultados de los distintos sistemas de personalización

Los dos primeros experimentos preliminares [Díaz *et al.* 00a, 01a] permitieron obtener conclusiones positivas sobre una versión inicial de aplicación de las técnicas de selección de contenidos presentadas en esta tesis. Sin embargo, también mostraron una serie de deficiencias tanto en su funcionamiento como en la evaluación efectuada.

Los resultados fueron altos en recall y precisión cuando se utilizaron sólo las secciones, fueron bajos en recall y precisión cuando se utilizaron sólo las categorías y fueron bajos en recall y altos en precisión cuando se utilizaron sólo las palabras clave.

Por otro lado, la evaluación cualitativa, basada en el cuestionario de evaluación rellenado por los usuarios, ofreció resultados aceptables en general, con una media de medio-alto para las distintas valoraciones efectuadas por los usuarios.

En estas dos primeras evaluaciones, el problema de la no-normalización de las distintas contribuciones de los distintos sistemas de referencia produce resultados que no reflejan el potencial de combinación de todos los métodos de clasificación. Además la representación de las categorías mediante las categorías de primer nivel de Yahoo! ofrece resultados pobres.

Por otro lado, se pusieron de manifiesto los problemas asociados a la evaluación de este tipo de sistemas. Por un lado, la necesidad de juicios de relevancia provenientes de usuarios reales y no de expertos humanos. Por otro lado, que los valores medios no se obtuvieran a partir de valores de recall iguales y que no se calcularan para varios puntos de recall para ver la evolución del sistema. Por otro, que los usuarios no dispusieran de la misma información

para juzgar todas las noticias. Y por último, que no se obtuvieran juicios de varios días de funcionamiento del sistema.

El tercer experimento preliminar [Acero *et al.* 01] permitió obtener conclusiones positivas sobre una versión inicial de aplicación de las técnicas de presentación de resultados. La evaluación mostró, en términos de precisión media para 11 niveles de recall, que los resúmenes genérico-personalizados funcionan mejor que los genéricos y peor que las noticias completas. Este estudio superficial se realizó sólo con 3 usuarios durante 1 día, pero permitió dar una idea del comportamiento de las técnicas empleadas.

En todo caso, se hizo patente después de estas primeras pruebas iniciales que el comportamiento del sistema es suficientemente complejo como para que las evaluaciones iniciales obtenidas sólo muestren parcialmente la efectividad real del mismo. En particular, la no-disponibilidad de juicios de relevancia sobre todas las noticias de todos los usuarios, y especialmente la falta de juicios para varios días distintos, limita los resultados obtenidos. Este último aspecto es especialmente importante por que el conjunto de noticias es diferente cada día y una selección efectiva para un día concreto no tiene porque serlo para todos los días.

Para solucionar estas deficiencias se desarrollaron dos colecciones de evaluación (colecciones 1.0 y 2.0) con varios usuarios durante varios días que permitieran obtener resultados más definitivos sobre la calidad de las técnicas propuestas. Además estas colecciones permiten, en primer lugar, probar el sistema con distintas configuraciones en sus parámetros para determinar la combinación de sistemas de referencia más interesante, y en segundo lugar, que otros prueben sus propuestas de personalización sobre el mismo marco de evaluación y facilitar así la comparación de resultados.

Las conclusiones obtenidas a partir de la aplicación de los sistemas de personalización 1.0 y 2.0 sobre las colecciones de evaluación 1.0 y 2.0 son similares: la combinación de todos los sistemas de referencia es mejor que la utilización de cualquier otra combinación, para las tareas de selección, adaptación y presentación. Sin embargo, en el sistema 1.0 no se utilizan las categorías y en el 2.0 sí se utilizan. De hecho, el último sistema se construyó para comprobar si la adición de este método de clasificación permitía corroborar las conclusiones obtenidas con el sistema 1.0 e incluso mejorar los resultados obtenidos.

Hay que destacar varios resultados en cuanto a las diferencias en los valores obtenidos en la evaluación del sistema de personalización 2.0 respecto al 1.0. Por un lado, que la selección utilizando secciones y palabras clave es similar en ambos sistemas (Tabla 7.1) a pesar de utilizar distintos usuarios, noticias y días para su evaluación. La combinación de secciones y palabras clave es ligeramente mejor en el sistema 2.0. Lo mismo ocurre con las secciones en solitario. Sin embargo, las palabras clave ofrecen ligeramente mejores resultados en el sistema 1.0. Esto es debido a que los problemas de especificidad y ambigüedad asociados a este método de selección en solitario se acentúan cuando hay más usuarios y más días. Además el incremento en las diferencias entre la combinación de secciones y palabras clave, y secciones y palabras clave en solitario permiten confirmar las conclusiones obtenidas con menos usuarios y durante menos días.

	Sistema 1.0		Sistema 2.0	
	nP	nR	nP	nR
SCP			0.588	0.685
SP	0.493	0.617	0.505	0.650
S	0.443	0.603	0.451	0.638
P	0.371	0.538	0.349	0.530

Tabla 7.1. Recall y precisión normalizados para el proceso de selección de contenidos en los sistemas de personalización 1.0 y 2.0.

En todo caso la inclusión de las categorías en el sistema 2.0 hace que las combinaciones de secciones, categorías y palabras clave produzca mejores resultados, en la selección de contenidos, que la combinación de secciones y palabras clave (Tabla 7.1). Por tanto, aunque las noticias y los usuarios son diferentes, se puede concluir que la utilización de las categorías incrementa la calidad de la información seleccionada por el sistema.

	Sistema 1.0		Sistema 2.0	
	nP	nR	nP	nR
S	0.443	0.603	0.451	0.638
C			0.497	0.605
P	0.371	0.538	0.349	0.530

Tabla 7.2. Comparación de la efectividad de los métodos de clasificación del modelo a largo plazo, en solitario, para la selección de contenidos en los sistemas de personalización 1.0 y 2.0.

Por otro lado, la comparación con los resultados obtenidos mediante los métodos de clasificación del modelo a largo plazo en solitario (Tabla 7.2) permite realizar las siguientes observaciones. Por un lado, las secciones siguen ofreciendo buenos resultados. Por otro lado, se observa la mejora obtenida para las categorías, que pasan a ser el mejor sistema en solitario en el sistema 2.0. Esto es debido a la mejora en la representación de las categorías conseguida añadiendo las páginas asociadas a las categorías de segundo nivel de Yahoo!. Por último, la efectividad obtenida con las palabras clave es debida a que los términos utilizados por los usuarios de los sistemas 1.0 y 2.0 estuvieron muy relacionados con intereses muy concretos, de hecho algunas de estas palabras aparecían en muy pocas noticias o en ninguna y otras se referían a temas tan generales que hacían que se diera especial valor a noticias en las que el usuario no estaba realmente interesado.

En cuanto al proceso de adaptación del modelo de usuario, se puede observar (Tabla 7.3) que las diferencias en la utilización del modelo a corto plazo frente a la no-utilización, combinado con cualquier combinación de métodos de clasificación del modelo a largo plazo, se incrementan en el sistema de personalización 2.0. También aumentan los valores absolutos de todas las combinaciones de métodos de clasificación con el modelo a corto plazo de un sistema a otro. Esto es debido al incremento en el número de días de evaluación (5 frente a 14), que permite que los efectos del modelo a corto plazo sean más patentes. Por lo tanto, se puede concluir, aunque las noticias y los usuarios son diferentes, que la utilización de combinaciones de métodos de clasificación del modelo a largo plazo junto con el corto plazo produce mejores resultados cuando los usuarios utilizan el sistema durante más días, es decir,

que el sistema se adapta mejor cuando se utiliza durante más tiempo. Este efecto también se puede observar en el incremento producido en la utilización del modelo a corto plazo en solitario.

	Sistema 1.0		Sistema 2.0	
	nP	nR	nP	nR
L(SCP)O			0.600	0.691
L(SP)O	0.531	0.624	0.568	0.669
L(SP)	0.489	0.617	0.505	0.650
L(S)O	0.515	0.604	0.535	0.652
L(S)	0.443	0.605	0.451	0.638
L(P)O	0.464	0.565	0.475	0.583
L(P)	0.365	0.534	0.349	0.530
O	0.399	0.511	0.421	0.545

Tabla 7.3. Recall y precisión normalizados para las distintas combinaciones de largo y corto plazo, para el proceso de adaptación del modelo de usuario en los sistemas de personalización 1.0 y 2.0.

En todo caso la inclusión de las categorías en el sistema 2.0 hace que las combinaciones de secciones, categorías y palabras clave, junto con el modelo a corto plazo, produzca mejores resultados, en la adaptación del modelo de usuario, que la combinación de secciones y palabras clave, junto con el modelo a corto plazo (Tabla 7.3). Por tanto, aunque las noticias y los usuarios son diferentes, se puede concluir que la utilización de las categorías incrementa la calidad de la adaptación del sistema.

En cuanto a la evolución en la mejora producida por la adaptación de contenidos a lo largo del tiempo, es decir, como va mejorando el recall y la precisión normalizada a lo largo de los días, se puede observar en la Tabla 7.4 que en el sistema 1.0 los resultados mejoran inicialmente a partir del tercer día debido al modelo a corto plazo y después fluctúan debido al cambio de noticias diario, produciéndose los mejores resultados el cuarto día. Los resultados altos obtenidos durante el primer día en el que no había modelo a corto plazo son debidos a que la definición del perfil de usuario se hace el día anterior al comienzo del funcionamiento del sistema, por tanto, los perfiles reflejan intereses que inevitablemente están relacionados con la información de la actualidad de la que disponen los usuarios ese mismo día.

	6-Mayo		7-Mayo		8-Mayo		9-Mayo		10-Mayo	
	nP	nR	nP	nR	nP	nR	nP	nR	nP	nR
L(SP)O	0.507	0.616	0.483	0.577	0.540	0.619	0.555	0.640	0.544	0.659

Tabla 7.4. Recall y precisión normalizados para la combinación de secciones y palabras clave del modelo a largo plazo, junto con el corto plazo, para cada día, para el sistema de personalización 1.0.

En el sistema 2.0 los resultados son similares (Tabla 7.5), es decir, los resultados del primer día comienzan siendo altos por similitud con el modelo a largo plazo definido por el usuario durante la semana anterior al comienzo de funcionamiento del sistema, el segundo día bajan debido a que esa similitud no es tan acusada, a partir del tercer día mejoran de nuevo los resultados y posteriormente fluctúan debido al cambio de noticias diario, produciéndose los mejores resultados el antepenúltimo día.

	nP	nR	nP	nR	nP	nR	nP	nR	nP	nR
	1-Dic		2-Dic		3-Dic		4-Dic		5-Dic	
L(SCP)O	0.633	0.726	0.420	0.531	0.550	0.665	0.590	0.680	0.641	0.725
O	0.243	0.500	0.318	0.489	0.410	0.543	0.409	0.544	0.410	0.532
			9-Dic		10-Dic		11-Dic		12-Dic	
L(SCP)O			0.603	0.689	0.598	0.700	0.611	0.704	0.623	0.710
O			0.385	0.510	0.438	0.548	0.424	0.538	0.467	0.572
	15-Dic		16-Dic		17-Dic		18-Dic		19-Dic	
L(SCP)O	0.688	0.766	0.615	0.701	0.707	0.788	0.689	0.751	0.684	0.748
O	0.478	0.602	0.481	0.588	0.473	0.571	0.460	0.560	0.449	0.550

Tabla 7.5. Recall y precisión normalizados para la combinación de secciones, categorías y palabras clave del modelo a largo plazo, junto con el corto plazo, para cada día, para el sistema de personalización 2.0.

El corto plazo en solitario se comporta de manera similar (Tabla 7.5) pero partiendo de unos resultados muy bajos que van aumentando a lo largo de los días con fluctuaciones debidas al cambio de noticias diario. El principal problema del modelo a corto plazo en solitario es que no dispone de un modelo de usuario inicial a partir del cual empezar a adaptarse. Esto tiene como consecuencia que las noticias del primer día respecto al modelo a corto plazo aparezcan en orden aleatorio, puesto que no hay respecto a lo que ordenarlas. Posteriormente, a partir de la realimentación se va adaptando poco a poco. Pero hay otra dificultad asociada a este modelo, el usuario sólo realimenta sobre las M noticias más relevantes, por lo tanto, puede ser que, inicialmente, las noticias sobre las que realimente no aporten nada positivo a su perfil.

En cuanto al proceso de presentación de resultados, se puede observar (Tabla 7.6) que los resultados obtenidos en la selección efectuada a partir de los resúmenes generados es claramente mejor en el sistema 2.0, esto es debido principalmente a la utilización de las categorías como método de clasificación para el modelo a largo plazo y al efecto del modelo a corto plazo durante más días de funcionamiento. En todo caso, los resultados para este sistema corroboran los obtenidos en el sistema 1.0 en el sentido de que los resúmenes personalizados son el mejor tipo de resumen. Sin embargo, entre los resúmenes genéricos y los resúmenes base se intercambian las posiciones en el ranking de todos los posibles resúmenes, esto es, en el sistema 2.0 son mejores los resúmenes base que los genéricos mientras que en el sistema 1.0 es al revés. Estas diferencias no son estadísticamente significativas, por lo que pueden ser debidas al cambio de usuarios y noticias. También se puede observar un pequeño incremento en la diferencia entre las noticias y los resúmenes personalizados, lo cual confirma que la selección efectuada a partir de las noticias completas es mejor que la que se obtiene con los resúmenes personalizados.

	Sistema 1.0		Sistema 2.0 (con categorías)		Sistema 2.0 (sin categorías)	
	nP	nR	nP	nR	nP	nR
N	0.526	0.622	0.603	0.694	0.563	0.669
Rp	0.523	0.621	0.593	0.686	0.569	0.673
Rgp	0.506	0.611	0.584	0.680	0.562	0.668
Rg	0.505	0.613	0.577	0.675	0.555	0.665
Rb	0.500	0.607	0.581	0.678	0.551	0.662

Tabla 7.6. Recall y precisión normalizados para los distintos tipos de resúmenes generados para el proceso de presentación de resultados en los sistemas de personalización 1.0 y 2.0, con categorías y sin categorías en el proceso de selección.

Por otro lado, se ha realizado un experimento para comprobar la influencia de las categorías en el proceso de selección de los resúmenes. Hay que tener en cuenta que las categorías no se utilizan en el proceso de generación de los resúmenes. Los resultados obtenidos son mejores cuando se utilizan las categorías que cuando no se utilizan (Tabla 7.6).

Sin embargo, hay que resaltar que los resúmenes personalizados son mejores que las noticias completas cuando no se utilizan las categorías en el proceso de selección de contenidos, esto es debido al proceso de selección de frases utilizado para la generación de los resúmenes personalizados, el cual se basa en la aparición de las palabras clave y los términos de realimentación del modelo de usuario. Si el proceso de selección de contenidos no utiliza categorías, y por tanto se basa sólo en secciones, palabras clave y términos de realimentación, la relevancia obtenida para los resúmenes personalizados es mayor porque precisamente estos resúmenes tienden a contener aquellas frases del documento que son más significativas con respecto a los dos últimos componentes utilizados en la selección de contenidos.

En todo caso los resultados del sistema 2.0 corroboran los del sistema 1.0 en cuanto a que los resúmenes personalizados son la mejor opción para la presentación de resultados, pero siempre peores que la noticia completa.

La diferencia en la evaluación del recall y precisión normalizados en los sistemas de personalización requiere una discusión más detallada. Los resultados menos claros respecto al recall normalizado pueden ser debidos al comportamiento de esta medida. Esto es, el recall normalizado es sensible a la clasificación del último documento relevante, mientras que la precisión normalizada es sensible a la clasificación del primer documento relevante. Los algoritmos propuestos mejoran la precisión porque hacen ascender en el ranking a los documentos más relevantes, pero también hacen descender a los menos relevantes. Esto puede ser debido a los juicios de relevancia de los usuarios, los cuales incluyen documentos muy relacionados con su perfil inicial y otros bastante alejados de él. Estos últimos documentos deberían ser detectados por la realimentación para corregir la tendencia. Sin embargo, dadas las peculiaridades del proceso de diseminación de la información, el usuario sólo puede suministrar realimentación sobre los M documentos más relevantes (los especificados como el número máximo de documentos que quiere recibir cada día). Esto implica que haya documentos, que si se mostraran al usuario, podrían ascender en el ranking gracias al proceso de realimentación, no sólo no ascienden sino que descienden en el ranking.

Sin embargo, es más interesante la precisión que el recall porque al usuario se le envían los documentos más relevantes de acuerdo al límite en el número de noticias que haya espe-

cificado en su perfil y el número de documentos relevantes seleccionados para un usuario es, en general, mayor que ese límite.

En todo caso, los resultados de la evaluación cualitativa realizada en el sistema de personalización 2.0 confirman las hipótesis comprobadas en la evaluación cuantitativa. Por un lado, que lo mejor es utilizar la combinación de métodos de clasificación para la selección y adaptación de contenidos, y por otro, que la utilización de los resúmenes ha sido satisfactoria para los usuarios puesto que los utilizaron habitualmente para decidir sobre la relevancia de las noticias y además los valoraron positivamente en cuanto a calidad, coherencia, claridad y adaptación. Las opiniones sobre las palabras clave fueron especialmente peculiares, los usuarios variaron mucho su opinión sobre ellas antes y después de usar el sistema, pero sin embargo siguieron prefiriéndolas a las secciones y a las categorías porque permiten definir sus intereses de manera muy concreta.

7.3. Comparación con estado del arte

La comparación explícita y directa con los resultados obtenidos en otros sistemas no ha sido posible porque no existía una colección de prueba, generada para algún otro sistema, con la que ejecutar el enfoque propuesto en esta tesis. Sin embargo, se han realizado comparaciones de distintas combinaciones de métodos de clasificación que representan distintas técnicas utilizadas en el estado del arte, de manera aislada en algunos trabajos y de manera conjunta en otros. Además la construcción de colecciones de evaluación permite que otras técnicas sean probadas y directamente comparadas entre sí, al estilo de lo que sucede con las colecciones estándar de recuperación de información.

En todo caso, se va a discutir a continuación las principales similitudes y diferencias con los trabajos más parecidos que aparecen en el estado del arte.

En algunos sistemas (p.ej.: [Billsus&Pazzani00]) se parte de un perfil vacío que se va actualizando por realimentación del usuario, esto puede llegar a frustrar a los usuarios si el sistema inicialmente selecciona muchos documentos irrelevantes. Hay que tener en cuenta que cuando un usuario utiliza por primera vez un sistema de personalización puede no tener muy claro cuales son sus necesidades de información. Por tanto, es más interesante partir de un modelo de usuario inicial que puede ser obtenido a partir de las respuestas a un cuestionario o puede ser introducido de manera explícita por el usuario. En cualquier caso conviene que la forma de introducir ese perfil inicial sea lo más clara y sencilla posible para evitar que suponga una carga de trabajo excesiva, pero sin que esa sencillez sea tan limitada que no permita reflejar adecuadamente los intereses de un usuario. El modelo a largo plazo propuesto en esta tesis permite una selección inicial de noticias interesantes para el usuario que además se pueden ir refinando mediante el modelo a corto a plazo obtenido a partir de la realimentación del usuario.

El hecho de realizar una selección global de todas las noticias de un día hace que no sea posible que el usuario vaya interactuando noticia a noticia y que la realimentación producida sobre una noticia afecte al resto de noticias del mismo día. Por ello, es necesario que el modelo de usuario almacene suficiente información como para poder realizar una selección razonable, aunque el usuario lleve poco tiempo utilizando el sistema. El modelo a largo plazo, rellenado por el usuario al darse de alta en el sistema, permite realizar una selección sin ningún esfuerzo inicial de realimentación por parte del usuario. El modelo a corto plazo permite que la realimentación sobre las noticias de un día sirva para la selección de las noticias de los días siguientes.

Los modelos de usuario utilizados en la bibliografía suelen estar basados en una única técnica para la representación de los intereses de los usuarios: términos [Chen&Sycara98], estereotipos [Ardissono *et al.* 01], redes semánticas [Asnicar *et al.* 97], redes neuronales [Shepherd *et al.* 02], etc., aunque existen algunos que combinan varias de ellas [Billsus&Pazzani00; Widyantoro01]. El modelo de usuario propuesto en este trabajo combina 4 técnicas diferentes que permiten una definición más completa de las necesidades de información de los usuarios.

El tipo de métricas utilizado para la evaluación, recall y precisión normalizados, no es la más habitual en los trabajos analizados, ya que estos se suelen basar simplemente en métricas que sólo tienen en cuenta las noticias hasta un cierto punto de recall fijado por el número de documentos que son juzgados por los usuarios. Estas métricas no tienen en cuenta todas las noticias presentes en el sistema y sobre todo no valoran el ranking final de los documentos, ya que valoran igualmente que haya 10 documentos relevantes entre 20, sean estos los más relevantes o los menos relevantes entre los 20.

Por otro lado, en [Billsus&Pazzani99a, 99b] se destaca la dificultad a la hora de evaluar ese tipo de sistemas debido a varias razones. En primer lugar no es posible aplicar las metodologías estándar de aprendizaje automático (validación cruzada) debido a que el orden cronológico de los ejemplos de entrenamiento no permite una selección aleatoria. Por otro lado hay que tener en cuenta que al medir el rendimiento diario se miden tanto los efectos de la realimentación del modelo como el efecto del cambio de noticias.

En cuanto a los resultados obtenidos respecto a la selección y adaptación de contenidos, con la combinación de todos los sistemas de referencia, se obtienen resultados similares a los obtenidos en otros sistemas. En [Billsus&Pazzani00] los resultados son parecidos en el sentido de que aumenta la efectividad (métrica F_1) durante los primeros días y luego fluctúa debido al cambio de noticias diario. Sin embargo, la efectividad inicial es muy baja porque no hay perfil inicial. Además, aunque el número de usuarios es superior (150 en la versión web y 185 en la versión PDA), el número de días de evaluación es inferior (3 días en la versión web y 10 en la versión PDA). En [Asnicar *et al.* 97] los resultados son similares a los anteriores (métrica precisión normalizada) pero se utilizan 4 usuarios durante 100 sesiones, 20 documentos por sesión. Además los documentos son resúmenes de informes técnicos no noticias de periódicos.

La mayoría de los criterios utilizados en la bibliografía para obtener los resultados de la evaluación son cuantitativos, es decir, asignan cantidades calculadas de diversas formas a las evaluaciones realizadas por el sistema en comparación con los juicios de relevancia determinados por los usuarios. Sin embargo, cada vez más se está optando por una evaluación cualitativa basada en opiniones de los usuarios, recogidas en cuestionarios, que muestran las impresiones de los usuarios sobre la utilización del sistema en diversos aspectos [Spink02; Beaulieu03]. En realidad, estas dos evaluaciones son complementarias y permiten visualizar el funcionamiento del sistema desde dos puntos de vista diferentes, desde el punto de vista del sistema y desde el punto de vista del usuario.

Una posible desventaja de la propuesta presentada es que no existe un mecanismo explícito para detectar novedades que pueden resultar de interés para los usuarios y que no estén reflejadas en su perfil. Sin embargo, el sistema posee cuatro sistemas de referencia a través de los cuales los usuarios pueden especificar sus intereses en eventos de características similares: secciones y categorías a las cuales la novedad puede pertenecer, palabras clave que pueden aparecer en ella o términos que se pueden haber obtenido de la realimentación sobre noticias similares.

Un método alternativo que podría permitir la detección de novedades sería el filtrado colaborativo [Claypool *et al.* 99]. Sin embargo, este método no es aplicable en el contexto en el cual funciona la propuesta de personalización presentada ya que el modo de envío está basado en las versiones diarias de los periódicos electrónicos. Esto hace que el proceso de personalización se realice a la vez para todos los usuarios y que el envío de los mensajes sea simultáneo para todos. Por tanto, los usuarios no pueden emitir juicios sobre las noticias que puedan servirle a otros para su propia personalización. A este problema se le denomina problema del primer evaluador, no existen juicios previos respecto a los cuales comparar.

En cuanto a la presentación de resultados, la mayoría de los sistemas de personalización utilizan opciones sencillas: títulos, títulos y primeras frases, títulos y frases donde aparecen las palabras de la consulta resaltadas. En todo caso, en los sistemas de personalización examinados en la bibliografía es poco común la posibilidad de mostrar al usuario un resumen de los documentos seleccionados como interesantes y mucho menos que estos resúmenes se adapten a los intereses de los usuarios descritos en sus modelos de usuario.

Las técnicas que se utilizan para generar los resúmenes se agrupan en tres tipos: técnicas de extracción de frases, técnicas basadas en relaciones discursivas y técnicas de abstracción. Se han utilizado las técnicas de selección y extracción de frases por su independencia del dominio y del idioma. Los posibles problemas de inconsistencia en el resumen resultante constituyen el principal inconveniente de esta aproximación, sin embargo, las opiniones mostradas por los usuarios en la evaluación cualitativa indican que su impresión es bastante positiva en cuanto a la coherencia y claridad de los resúmenes.

En principio, la evaluación de la presentación de resultados se debería realizar a partir de una evaluación cualitativa donde los usuarios emitieran sus opiniones acerca de la calidad de esos contenidos generados. Lógicamente esa calidad será mayor cuanto mejor ayuden al usuario a resolver sus necesidades de información. En todo caso, también se puede medir cuantitativamente la presentación teniendo en cuenta la utilización de los resultados presentados para resolver otra tarea de acceso a la información, o comparándolo con una presentación de resultados "ideal". En particular, cuando la presentación se realiza utilizando resúmenes lo que se intenta determinar cuantitativamente es cómo de adecuado o de útil es un resumen con respecto al texto completo [Hahn&Mani00]. Existen distintos enfoques que pueden clasificarse, según [Sparck-Jones&Galliers96], en directos (o intrínsecos) e indirectos (o extrínsecos). Los primeros se basan en el análisis directo del resumen a través de alguna medida que permita establecer su calidad. Los segundos basan sus juicios en función de su utilidad para realizar otra tarea.

Aunque el método de evaluación más utilizado para medir la relevancia del contenido de un resumen automático es la comparación con un resumen "ideal" construido manualmente, los problemas de concordancia entre jueces a la hora de elegirlo, el hecho de que un resumen sólo se considere bueno si se parece al resumen ideal y la dificultad añadida de que en este caso los jueces tendrían que construir el resumen más adecuado para cada usuario para cada noticia, hacen que este tipo de evaluación no sea la más adecuada en el marco de la generación de resúmenes personalizados presentada en este trabajo.

Sin embargo, la evaluación indirecta sin jueces permite una evaluación mucho menos cara y mucho más extrapolable a grandes *corpora* de texto y a diversas técnicas de generación de resúmenes con diversas tasas de compresión.

Esta técnica se ha utilizado en trabajos similares en los que se ha medido el efecto en la recuperación de información de la utilización de distintos tipos de resúmenes en lugar de la noticia completa. En [Maña *et al.* 98, 99, 00] se utilizaron 5000 documentos y 50 consultas y los resultados (métrica precisión media en 11 niveles de recall) mostraron que los resúmenes

adaptados a la consulta mejoran la precisión obtenida con las primeras frases de los textos e igualmente con los resúmenes genéricos (heurísticas de palabras temáticas, localización y título). También se observó que las noticias completas mejoran de manera no significativa a los resúmenes adaptados. Hay que tener en cuenta que los resúmenes adaptados a la consulta son en realidad equivalentes a los resúmenes genérico-personalizados presentados en este trabajo y que no se evalúa el efecto de los resúmenes personalizados porque se supone que su contenido no va guardar fidelidad con el contenido del texto original. En [Nomoto&Matsumoto01] se realiza un estudio similar con 5000 artículos de un periódico japonés y 50 consultas. Los resultados son similares aunque la métrica utilizada es diferente (F_1). También en [Tombros&Sanderson98] se compara el segmento inicial con resúmenes orientados a la consulta. Se utilizan 50 consultas TREC y 50 documentos por consulta y se mide la precisión, el recall, la velocidad en el proceso de decisión, las veces que el usuario accede al texto completo y la opinión de los usuarios sobre la calidad de la información (resumen o segmento inicial) suministrada. Los resultados muestran que los resúmenes orientados al usuario mejoran la eficacia de los usuarios en tarea de recuperación respecto a la utilización de los segmentos iniciales.

El dominio de los periódicos electrónicos ofrece un buen ejemplo donde la estructura de los documentos da muy buenas pistas de la relevancia de su contenido (las noticias tienen habitualmente la forma de pirámides invertidas en términos de relevancia: la información más relevante está al comienzo, y la relevancia va decayendo conforme se avanza a lo largo del documento). Esto hace que las técnicas simples de generación de resúmenes ofrezcan buenos resultados en términos de generación de resúmenes indicativos, permitiendo que el usuario obtenga una idea de cual es el tema de la noticia, aunque dicho método podría ser muy dependiente del dominio y podría no trabajar bien en otros dominios.

Los métodos propuestos aquí aseguran la selección de la información relevante en términos de necesidades de información, funcionando eficientemente para la generación de resúmenes personalizados, suministrando a cada usuario un extracto de cada documento con los contenidos específicos que están más relacionados con sus intereses, de manera independiente del dominio.

7.4. Extrapolación a un ámbito multilingüe

Este apartado va a tratar sobre un primer intento de extrapolación del sistema de personalización propuesto a un ámbito multilingüe. La idea es tener dos fuentes de documentos Web (dos periódicos digitales) en dos idiomas distintos, un modelo de usuario que sirva para reflejar los intereses de los usuarios en los dos idiomas y un sistema de personalización que seleccione las noticias más relevantes en ambos idiomas y las presente de manera personalizada, cada una en su idioma original. Además se debe poder realimentar el sistema y esta realimentación debe producir la adaptación del modelo de usuario, independientemente del idioma de la noticia sobre la que se realimente.

7.4.1. Introducción

Un sistema de personalización multilingüe está basado en las mismas 3 fases indicadas para los sistemas monolingües, esto es, selección, adaptación y presentación. La dificultad es que los documentos de entrada y los modelos de usuario pueden estar en varios idiomas y por lo tanto que los procesos tienen que adaptarse al idioma correspondiente.

Las técnicas empleadas en personalización multilingüe son similares a las que se utilizan en recuperación de información multilingüe (CLIR, *Cross Lingual Information Retrieval*) [Grefenstette98a]. Esta área de investigación está dedicada a la tarea de filtrado, selección y clasificación de documentos que pueden ser relevantes a una consulta expresada en un idioma diferente al de los documentos. Las posibilidades básicas son traducir los documentos, traducir la consulta y utilizar una representación intermedia.

La propuesta más común es la que se basa en la traducción de la consulta (p.ej.: todos los experimentos multilingües en TREC-10 [Gey&Oard01]). Esto es debido a que el tamaño de la consulta es mucho menor que el de los documentos. Los principales problemas de este enfoque son: como expresar un término de un determinado idioma en otro idioma diferente, determinar qué traducciones son las más adecuadas, asignar pesos a las diferentes traducciones. Para solucionarlos se pueden utilizar recursos electrónicos como diccionarios bilingües, *corpora*, programas de traducción automática o tesauros.

En general no va a haber una única traducción por término, sino que habrá varias posibles que dependerán del contexto en el que se utiliza el término. Por tanto, habrá que decidir cual de las posibles traducciones utilizar en la consulta destino. Existen distintas opciones [Díaz *et al.* 02; LópezOstenero *et al.* 04] dependiendo del recurso electrónico utilizado. La traducción de los documentos en lugar de las consultas presenta ventajas en cuanto a la posible calidad de la traducción, pero grandes desventajas en carga computacional [Oard98]. La opción de la representación intermedia independiente del idioma consiste en la traducción de consultas y documentos a esta representación y la realización de la búsqueda en esa nueva representación. Una posibilidad es utilizar EuroWorNet como representación intermedia [Gilarranz *et al.* 97].

Las tres posibilidades básicas de la recuperación de información multilingüe tienen sus equivalentes cuando se trata de personalizar contenidos: la primera posibilidad consiste en tener un único modelo por usuario, en el idioma seleccionado por el usuario, y traducir los documentos que estén en otro idioma al idioma del usuario. La segunda posibilidad consiste en tener un modelo de usuario por idioma, de tal forma que el usuario introduce el modelo en su idioma y el sistema traduce el modelo al resto de los idiomas que maneje el sistema de personalización. La tercera consiste en encontrar una representación independiente del idioma tanto para los documentos como para el modelo de usuario.

Una vez realizadas las traducciones, de documentos o de modelos, el proceso de selección se puede realizar en un sólo idioma, si se utiliza la primera posibilidad, o en paralelo para cada uno de los idiomas manejados, si se elige la segunda. En este último caso, hay que mezclar los resultados de varios idiomas teniendo en cuenta la relevancia asignada a cada uno de los documentos, estén en el idioma que estén. En todo caso las técnicas empleadas cuando se trabaja en un sólo idioma pueden ser cualesquiera de las descritas en los apartados anteriores. Por último, si se elige la tercera posibilidad, el proceso de selección estará basado en la representación independiente del idioma.

7.4.2. Estado del arte

En la actualidad hay pocos trabajos que muestren personalización multilingüe, en ITEM [Verdejo00] se propone un buscador multilingüe, español-catalán-vasco-inglés basado en conceptos [Gonzalo *et al.* 99] en lugar de términos. Este buscador realiza análisis morfológico, etiquetado sintáctico y desambiguación para generar representaciones independientes del idioma de documentos y consultas. Los descriptores de indexación corresponden al índice

InterLingua (superconjunto de todos los conceptos que aparecen en todos los idiomas) de la base de datos léxica EuroWordNet/ITEM, que conecta las 4 wordnets en cada idioma.

En SiteIF [Magnini&Strapparava01] se presenta un sistema de filtrado de noticias en inglés e italiano. Tanto los documentos como el modelo de usuario se representan utilizando conceptos independientes del idioma [Gonzalo *et al.* 98]. Para ello se utiliza la base de datos léxica multilingüe inglés-italiano MultiWordNet [Artale *et al.* 97], donde los sentidos están alineados. Los conceptos se localizan a partir de los términos que aparezcan en los documentos en cada uno de los idiomas. Para elegir el sentido adecuado se utiliza una técnica llamada desambiguación de palabras de dominio [Magnini&Strapparava00], que se basa en la asignación de etiquetas de dominio a cada uno de los sentidos asociados a una palabra.

En OmniPaper [González *et al.* 02] se pretende ofrecer un medio de acceso personalizado y unificado a las noticias de periódicos europeos. El objetivo es crear una capa de navegación y enlace por encima de recursos de información distribuidos que permita buscar noticias en todos los recursos manejados con una única consulta, expresada en un único idioma. El manejo de varios idiomas está basado en la existencia de una taxonomía multilingüe que almacena conceptos con sus términos asociados en cada uno de los idiomas, además de relaciones semánticas entre los conceptos. Cada uno de los recursos distribuidos se enlaza con dicha taxonomía a través de módulos de traducción. El nivel de conocimiento global también almacena la representación de los usuarios y su posterior adaptación.

7.4.3. Personalización multilingüe de contenidos Web

El sistema de personalización multilingüe (español-inglés) de contenidos Web que se propone se basa en un sistema de filtrado de información basado en un modelo de usuario en dos idiomas, es decir, se tendrá un modelo de usuario en español y otro en inglés, pero equivalentes [Díaz01]. En este caso la información que va a ser filtrada son documentos Web en español y en inglés y lo que se va a ofrecer a los usuarios, el resultado del filtrado, es un único documento Web con los contenidos adecuados según su modelo de usuario, de tal forma que los documentos en inglés serán seleccionados por el modelo en inglés y los documentos en español por el modelo en español. De hecho, el diseño de los procesos de personalización se ha realizado de forma que sean fácilmente aplicables a distintos idiomas.

Una opción sería que cada usuario rellenara los dos modelos (categorías propias, categorías generales y palabras clave) como si fueran independientes y que el sistema filtrara los documentos en cada uno de los idiomas según su modelo. En realidad sería duplicar el sistema monolingüe cambiando únicamente el extractor de raíces y la lista de palabras vacías. Lógicamente esto haría que la adaptación y la presentación también fueran independientes y los procesos con los documentos en un idioma no influyeran ni se vieran influidos por los del otro idioma. Sin embargo, este funcionamiento no es muy operativo puesto que la realimentación que pueda introducir un usuario sobre un documento, esté en el idioma que esté, debería afectar a la selección de nuevos documentos estén en el idioma que estén. Es decir, que sólo debe haber un modelo de usuario equivalente para los dos idiomas. Para ello, lo que se realiza es una traducción de un modelo a otro para que la información sea la misma en ambos idiomas.

Inicialmente, el usuario elige el idioma con el que quiere trabajar, de tal forma que todas las interacciones con el sistema se realicen en dicho idioma. Posteriormente selecciona las distintas partes de su modelo de usuario: información personal, información sobre el forma-

to de la información recibida e información específica sobre los intereses del usuario según varios sistemas de referencia que serán los que se utilicen para realizar la personalización.

A la hora de seleccionar sus intereses, el usuario tiene que introducir las palabras clave en el idioma seleccionado. La elección sobre las categorías generales se tiene que realizar sobre el sistema de categorías de primer nivel de Yahoo! del idioma seleccionado. En particular, al haber seleccionado el inglés como segundo idioma las categorías generales en el modelo en inglés corresponden a las de Yahoo! Estados Unidos: Arts, Business and Economy, Computers and Internet, News and Media, Recreation, Reference, Education, Regional, Entertainment, Science, Government, Social Science, Health, Society and Culture.

Posteriormente el sistema traduce el modelo en el idioma seleccionado al modelo en el otro idioma traduciendo las palabras clave y las categorías. La traducción de las palabras clave se realiza mediante un traductor automático [Giráldez *et al.* 02] que establece correspondencias entre los términos en ambos idiomas de tal forma que si existen varias traducciones para un término se elige la primera posibilidad, supuesta la más habitual. Por otro lado, la elección sobre las categorías es independiente del idioma porque existe una correspondencia uno a uno entre el sistema de categorías de primer nivel de la versión española de Yahoo! y el de la versión en inglés. Por tanto, la traducción de las categorías simplemente es copiar los pesos elegidos en la versión seleccionada a la otra versión en el otro idioma. El sistema de clasificación propuesto, las categorías de Yahoo!, está disponible en muchos idiomas con las mismas categorías, lo cual permite construir la representación de las categorías a partir de las páginas asociadas a cada categoría en su propio idioma

En cuanto a las categorías propias, en general, no se puede establecer una correspondencia como la establecida entre las categorías generales. Por ejemplo, que un usuario seleccione la sección de internacional en un periódico electrónico en un idioma no quiere decir que esté interesado en todas las noticias de nacional del otro idioma. Por ello, el usuario tiene que seleccionar las categorías propias en ambos idiomas en las que está interesado.

Por tanto, el usuario indica sus intereses sobre las categorías propias en ambos idiomas, y sobre las categorías y las palabras clave en el idioma que haya seleccionado. Posteriormente, el sistema genera el modelo en el otro idioma traduciendo las categorías y las palabras clave.

Para obtener el modelo a corto plazo mediante la adaptación del modelo de usuario se aplica el proceso descrito en el apartado 3.3.2.1 independientemente del idioma en el que se encuentre el documento sobre el que se efectúe la realimentación. Se obtienen una serie de términos de realimentación que pueden estar en idiomas diferentes. Se actualizan los modelos a corto plazo en cada idioma utilizando el término de realimentación en el mismo idioma obtenido en la realimentación o traducido, si el idioma del modelo a corto plazo no coincide.

La ventaja de modelar a los usuarios de esta manera es que el sistema es fácilmente adaptable a otros idiomas y extensible a varios más. Lo único que habría que cambiar o añadir serían los procesos que son dependientes del idioma, esto es, el extractor de raíces, la lista de palabras vacías y el traductor. También habría que cambiar las categorías de primer nivel de Yahoo!, pero siempre existe correspondencia entre los sistemas de categorías de varios idiomas. Además habría que cambiar o añadir las categorías propias en cada uno de los idiomas.

La representación de los documentos se realiza utilizando el MEV en cada uno de los idiomas, es decir, se representan los documentos en español por un lado, y los documentos en inglés por otro. Para cada representación se utilizan los procesos dependientes del idioma correspondientes, esto es, la lista de palabras vacías y el extractor de raíces.

Las secciones se representan en matrices de pesos asociados a las secciones de cada idioma. Las categorías se representan utilizando los términos que aparecen en las páginas in-

dexadas bajo las categorías de primer nivel de Yahoo! en el idioma correspondiente. Las palabras clave y los términos de realimentación se representan también como vectores de pesos de términos en cada uno de los idiomas que maneja el sistema.

Lo que se realiza es una selección de contenidos en cada idioma, esto es, los documentos en español son procesadas por el modelo en español de la misma manera explicada en el apartado 3.4, y lo mismo ocurre con los documentos en inglés y el modelo en inglés.

Se obtienen dos rankings a partir de la ecuación (3.10), uno en cada idioma, que son mezclados en un solo ranking. Para realizar la mezcla se supone que los valores obtenidos en ambos idiomas son directamente comparables, por lo que no hay que hacer ningún tipo de ajuste. Los documentos con mayor relevancia son los que se envían al usuario, cada uno en su idioma original.

Finalmente se generan los resúmenes de los documentos seleccionados en su idioma original utilizando las heurísticas descritas en el apartado 3.5, ya que tanto la heurística de posición, como la de palabras clave como la de personalización, son independientes del idioma. Para la heurística de personalización se utiliza el modelo de usuario en el idioma correspondiente al documento que se va a resumir.

7.4.4. Minicolección multilingüe

Se ha generado una minicolección "multilingüe" para evaluar un prototipo del sistema de personalización multilingüe. Contiene 13 usuarios durante 6 días (del 18 al 24 de Septiembre de 2001) y utiliza 8 secciones del periódico español ABC y 6 del periódico inglés The Guardian, que almacenan aproximadamente 200 noticias por día (más o menos la mitad por idioma). Sólo almacena los intereses del perfil en un idioma, excepto las secciones que lo almacena en los dos por no existir correspondencia directa. Los juicios de relevancia corresponden a un único día y son generados por un experto humano para los dos idiomas, además no se almacena información sobre juicios de realimentación.

Se ha dispuesto de 13 modelos de usuario distintos. Los modelos de usuario iniciales fueron construidos por los usuarios el día anterior al comienzo del experimento. Estos perfiles iniciales contienen información sobre los intereses a largo plazo del usuario, esto es, idioma, secciones del periódico español, secciones del periódico inglés, categorías y palabras clave. El perfil de usuario se maneja en el idioma elegido por el usuario, traduciéndose las palabras clave al otro idioma y estableciéndose la correspondencia entre las categorías en el idioma del usuario y las del otro idioma.

El tercer paso de la construcción de la colección es la obtención de los juicios de relevancia de cada uno de los usuarios. Se eligió el último día para que el usuario realizara una evaluación exhaustiva del sistema, esto es, debía indicar cuántas noticias de las que recibía eran relevantes y cuántas de las que no recibía eran relevantes. Sin embargo, sólo se indicó a los usuarios que mostraran el número de noticias relevantes y no cuáles eran relevantes. Por lo tanto, esta información no se pudo utilizar para construir la colección de evaluación.

Los juicios de relevancia se obtuvieron examinando cada uno de los modelos de usuario con respecto a todas las noticias de ese día (207) y determinando cuáles eran las relevantes. Este examen no lo hacen cada uno de los usuarios sino que lo realiza un único evaluador para todos los modelos. El número medio de noticias relevantes por usuario fue 121.4, sobre 207 posibles.

Por último, resaltar que aunque el sistema permitía la realimentación por parte de los usuarios, esta información no fue recogida en el momento de la evaluación por lo que la colección no dispone de la misma.

7.4.5. Evaluación

Puesto que en el caso multilingüe existen los mismos tres procesos de personalización, esto es, selección, adaptación y presentación, se puede aplicar la misma metodología de evaluación presentada para el caso monolingüe. La principal diferencia estriba en el conjunto de noticias respecto al cual se evalúa el sistema, ya que en el caso de la personalización multilingüe hay que tener en cuenta las noticias en ambos idiomas, eso hace que en los rankings de noticias, haya noticias de los dos idiomas mezcladas pero ordenadas con respecto al modelo de cada usuario.

En esta evaluación el usuario podía elegir cuantas noticias iba a contener el mensaje que iba a recibir, es decir, no se le mandaba un número fijo de noticias, sino el número de noticias que él elegía.

Se realizó una evaluación cualitativa, basada en impresiones de los usuarios recogidos en un formulario (Apéndice I) que tenía que ser rellenado por los usuarios el último día de la evaluación. Sólo 6 usuarios lo cumplimentaron. También se realizó una evaluación cuantitativa el último día de funcionamiento del sistema. Esta evaluación fue realizada por una única persona examinando los logs generados por el sistema de los 13 usuarios (colección de evaluación multilingüe).

Los resultados obtenidos son similares a los obtenidos en el primer experimento preliminar, esto es, resultados altos en precisión tanto en la evaluación cuantitativa como en la cualitativa y resultados altos en recall en la cuantitativa pero bajos en la cualitativa.

El sistema multilingüe evaluado adolecía del problema de la no-normalización de las distintas contribuciones de los distintos sistemas de referencia produciendo resultados que no reflejan el potencial de combinación de todos los métodos de clasificación. Además la representación de las categorías mediante las categorías de primer nivel de Yahoo! ofrece resultados pobres para este sistema de referencia.

Por otro lado, se pusieron de manifiesto los mismos problemas detectados en la evaluación de los experimentos preliminares de la versión monolingüe. Por un lado, la necesidad de juicios de relevancia provenientes de usuarios reales y no de expertos humanos. Por otro lado, que los valores medios no se obtuvieran a partir de valores de recall iguales y que no se calcularan para varios puntos de recall para ver la evolución del sistema. Y por último, que no se obtuvieran juicios de varios días de funcionamiento del sistema.

En todo caso, la evaluación cualitativa ha mostrado una cierta satisfacción por parte de los usuarios a pesar de ser una primera versión de sistema multilingüe [García *et al.* 02; Giráldez *et al.* 02]. Es necesario refinar aún más el sistema y sobre todo realizar una evaluación mucho más exhaustiva con más usuarios y durante más días, al estilo de la realizada para el sistema de personalización 2.0.

7.5. Resumen y conclusiones del capítulo

En este capítulo se ha presentado la comparación de los resultados obtenidos en las distintas propuestas de personalización, así como las similitudes y diferencias con los trabajos similares que aparecen en el estado del arte.

Los resultados obtenidos en los experimentos preliminares mostraron una serie de problemas asociados a la evaluación de este tipo de sistemas: normalización de resultados, representación de las categorías, necesidad de juicios de relevancia de usuarios reales, más usuarios y sobre todo más días de evaluación.

Los resultados obtenidos con los sistemas de personalización 1.0 y 2.0 mostraron que la combinación de sistemas de referencia es mejor que cualquier otra combinación, para las tareas de selección, adaptación y presentación.

Para el proceso de selección, los resultados fueron ligeramente mejores en el sistema 2.0 excepto para las palabras clave, donde se acentuó el problema de su utilización en solitario. La utilización de las categorías en el sistema 2.0 incrementó la calidad de la información seleccionada por el sistema, debido a la mejor representación utilizada para las mismas.

En el proceso de adaptación, los resultados mejoraron en el segundo sistema debido al incremento en el número de días de evaluación, es decir, que el sistema se adapta mejor cuando se utiliza durante más tiempo. Por otro lado, la utilización de las categorías en el sistema 2.0 incrementó la calidad en la adaptación del sistema.

Los resultados obtenidos a lo largo del tiempo presentan, en los sistemas 1.0 y 2.0, un decremento inicial, un posterior incremento y una fluctuación posterior debido al cambio de noticias diario. Los buenos resultados iniciales se producen porque los perfiles de usuario se definen basándose en información sobre la actualidad del día antes de comienzo de funcionamiento del sistema. Las mejoras y fluctuaciones posteriores son por efecto del modelo a corto plazo y del cambio de noticias diario.

En cuanto al proceso de presentación de resultados, se confirman en el sistema 2.0 los resultados del sistema 1.0 en cuanto a la mejora producida con los resúmenes personalizados respecto a los otros tipos de resúmenes, aunque la presencia de las categorías y del modelo a corto plazo durante más días incrementan los resultados obtenidos para la selección de todos los tipos de resúmenes.

Los resultados de la evaluación cualitativa confirman los resultados obtenidos en la evaluación cuantitativa en cuanto a utilización de varios sistemas de referencia y en cuanto a utilización de los resúmenes personalizados para la presentación de resultados.

En cuanto a la comparación con el estado del arte, resaltar que, aunque no se pueden realizar comparaciones directas por falta de colecciones de prueba, las distintas combinaciones de técnicas utilizadas representan distintos enfoques utilizados en el estado del arte. Además la construcción de colecciones de evaluación permite que otros puedan realizar estas comparaciones directamente.

En cuanto a las diferencias con otros sistemas cabe destacar la utilización del modelo a largo plazo como modelo inicial del cual carecen algunos sistemas, la utilización de 4 sistemas de referencia para la selección de contenidos y el uso de métricas que tienen en cuenta el ranking de los resultados.

En cuanto a resultados obtenidos, se han observado comportamientos parecidos en otros sistemas de personalización similares con mejoras iniciales y fluctuaciones posteriores, aunque estos sistemas presentaban malos resultados al principio por la ausencia de modelo ini-

cial y además el número de usuarios y días era inferior a los presentados en el sistema de personalización 2.0.

Se han utilizado las técnicas de selección y extracción de frases para la generación de resúmenes por su independencia del dominio y del idioma. En cuanto a la evaluación se ha optado por una evaluación indirecta basada en el efecto de la selección de contenidos sobre los distintos tipos de resúmenes, debido a su facilidad y extensibilidad, además la evaluación directa resulta inadecuada debido a los problemas de concordancia de jueces y a que en este caso los resúmenes son personalizados, y por tanto hay uno distinto por noticia y por usuario.

Los resultados obtenidos en trabajos similares en los que se ha medido el efecto de los resúmenes adaptados a la consulta de un usuario frente a segmentos iniciales en tareas de recuperación de información, confirman los resultados obtenidos en este tesis doctoral.

Se ha presentado la personalización multilingüe como generalización de los procesos descritos en un ámbito monolingüe. La idea es utilizar dos modelos de usuario, uno en cada uno de los idiomas que se manejan, y realizar los procesos de selección, adaptación y presentación con los documentos que estén en el mismo idioma que el modelo, y posteriormente mezclar los resultados directamente. De hecho, el diseño de los procesos de personalización se ha realizado de forma que sean fácilmente aplicables a distintos idiomas.

La ventaja de modelar a los usuarios de esta manera es que el sistema es fácilmente adaptable a otros idiomas y extensible a varios más. Lo único que habría que cambiar o añadir serían los procesos que son dependientes del idioma, esto es, el extractor de raíces, la lista de palabras vacías y el traductor. Además habría que cambiar o añadir las categorías propias en cada uno de los idiomas.

La personalización multilingüe ha ofrecido resultados similares a los obtenidos en el primer sistema de personalización aunque es necesario una evaluación mucho más detallada para poder obtener conclusiones más definitivas. En todo caso, los usuarios han mostrado en la evaluación cualitativa una cierta satisfacción con la utilización del sistema.

Capítulo 8
CONCLUSIONES

8.1. Principales aportaciones

Se ha presentado un modelo de usuario que representa de manera separada los intereses a largo plazo junto a las necesidades a corto plazo generadas a partir de la interacción con las noticias recibidas. Los usuarios pueden expresar sus preferencias a largo plazo en función de un sistema de clasificación dependiente del dominio (secciones) y a partir de información independiente del dominio basada en el contenido de los documentos. Esta información se subdivide a su vez en información procedente de un sistema de clasificación independiente del dominio (categorías de Yahoo!) y de un conjunto de palabras clave. Esta representación de los intereses de un usuario funciona como si fuera la definición de un estereotipo evitando comenzar con un modelo vacío que tiene que ser construido mediante la realimentación inicial del usuario sobre un conjunto de documentos. La utilización de un modelo inicial vacío puede resultar frustrante para los usuarios si al principio reciben muchos documentos que no son relevantes para sus intereses. Los intereses a largo plazo permiten una selección inicial de noticias interesantes para el usuario que además se pueden ir refinando mediante el modelo a corto a plazo obtenido a partir de la realimentación del usuario.

Es evidente que la utilización de secciones, categorías o palabras clave sirve para distintos propósitos. Las secciones son más adecuadas cuando el usuario quiere identificar un interés general sobre todos aquellos documentos que pertenezcan a una determinada sección prefijada por los editores del periódico. Las palabras clave tienen más utilidad cuando el usuario quiere definir un interés más concreto. Las categorías permiten al usuario una selección general más intuitiva más allá de la selección rígida asociada a las secciones. Para una aplicación de filtrado de documentos, utilizar una combinación de sistemas de referencia permite a los usuarios la definición de sus intereses desde distintos puntos de vista.

En general, el modelado a largo y corto plazo suponen diferentes alternativas que dan servicio a diferentes necesidades. El modelado a largo plazo es más adecuado para aplicaciones de filtrado de información porque no requiere realimentación del usuario para funcionar eficientemente y puede capturar los intereses reflejados por el usuario en su perfil. Por otro lado, no puede seguir la pista de los cambios en los intereses del usuario que se produzcan frecuentemente. El modelado a corto plazo presenta claras ventajas en casos en los que no hay un conjunto de intereses establecido sino una cambiante necesidad en el filtrado de la información debido a un cambio frecuente en los intereses del usuario. Sin embargo, requiere una considerable cantidad de realimentación antes de que el modelo pueda funcionar eficientemente, y el proceso de entrenamiento puede ser largo. La combinación de ambos modelos presenta ventajas en los casos en los que las necesidades de ambos tipos aparecen en el mismo contexto: hay unos intereses a largo plazo que considerar, así como hay que tener en cuenta los cambios recientes que se produzcan.

La propuesta de personalización presentada incluye un proceso de presentación de resultados basado en la construcción de resúmenes personalizados para cada noticia para cada usuario. Este tipo de presentación ha permitido a los usuarios decidir sobre la relevancia de las noticias recibidas sin necesidad de inspeccionar el texto completo.

Por otro lado, se ha presentado una evaluación sistemática de diferentes alternativas a la hora de realizar la personalización de contenidos. Se han construido una serie de colecciones de evaluación no sólo para permitir evaluar las distintas versiones de sistemas de personalización propuestos, sino también para que otras propuestas puedan ser evaluadas y comparadas con las presentadas en esta tesis.

La evaluación cuantitativa ha mostrado que la utilización de combinaciones lineales de todos los sistemas de referencia del modelo de usuario ofrece mejores resultados, tanto en la selección como la adaptación de contenidos, que la utilización de cualquier otra combinación de los mismos. Con respecto, a la generación de resúmenes, los resúmenes personalizados funcionan mejor que cualquier otro tipo de resumen, constituyendo una alternativa interesante, mucho mejor que las primeras líneas de la noticia. Además los resultados sólo ligeramente peores respecto a las noticias completas dejan abierta la posibilidad de utilizar los resúmenes en lugar de las noticias cuando sea aceptable una pequeña pérdida de información.

La evaluación cualitativa ha permitido contrastar los resultados obtenidos en la evaluación cuantitativa a través de las opiniones de los usuarios. Los cuestionarios de evaluación han sido desarrollados para permitir a los usuarios emitir sus juicios sobre el sistema de la manera más completa posible.

Los distintos procesos han sido parametrizados para permitir su adaptación a distintas configuraciones de personalización. En este trabajo, en particular, ha sido fijado el valor de algunos de estos parámetros (Tabla 3.2) para permitir el estudio del resto de ellos (Tabla 3.3). Algunos de los valores de los parámetros fijos han sido seleccionados basándose en estudios similares encontrados en la bibliografía, y otros han sido elegidos como valores razonables. También ha sido discutida la influencia de estas decisiones en aspectos relevantes de la personalización, como por ejemplo la utilización de distintas métricas o el efecto de la realimentación debido al parámetro M que gobierna el número de noticias que recibe un usuario cada día. En todo caso, se podrían variar también los valores de los parámetros fijos si se decidiera experimentar con ellos.

Las técnicas empleadas pueden ser fácilmente aplicadas a otros dominios siempre y cuando se disponga de una clasificación dependiente del dominio y de descripciones textuales de los elementos a seleccionar. Adicionalmente, el documento Web generado para cada usuario puede ser portado a tecnologías WAP/PDA simplemente cambiando el lenguaje de marcado (WML/HTML simplificado) [Acero&Alcojor01].

También hay que resaltar que, aunque las técnicas de personalización presentadas en este trabajo se han aplicado utilizando el español, se pueden utilizar directamente con cualquier otro idioma simplemente cambiando el conjunto de palabras vacías y el extractor de raíces.

La extrapolación a un ámbito multilingüe ha mostrado una evaluación positiva en la valoración de un primer prototipo de sistema, aunque es necesaria una evaluación mucho más detallada para obtener conclusiones más definitivas.

8.2. Trabajo futuro

Existen varias líneas de trabajo que dan continuidad al presentado en esta tesis. Por un lado, las colecciones de evaluación generadas y la parametrización del sistema van a permitir explorar otras muchas combinaciones, tanto de los parámetros variables como de los parámetros fijos, que van a permitir implementar distintas variaciones en las técnicas de selección, adaptación y presentación. También se van a probar otras propuestas diferentes de los distintos procesos de personalización que podrán ser comparadas directamente con los resultados obtenidos en esta tesis.

Otra posibilidad que se va a utilizar en nuestro sistema de personalización es la de la adaptación en la representación de las categorías a lo largo del tiempo. Para ello se va a utilizar la información obtenida de la categorización diaria para mejorar la representación de las categorías. Se utilizará la representación de aquellos documentos que obtengan una mayor relevancia en la categorización con respecto a cada categoría. Con estos documentos se realimentará la representación inicial que irá variando ahora cada día.

También se va a realizar una evaluación de los resultados obtenidos agrupando a los usuarios según sus tipos de intereses. De esta manera se podrá detectar si alguno de los métodos de clasificación es especialmente significativo para alguno de estos grupos y se podrá utilizar esta información para que el sistema sea capaz de ajustar sus parámetros según el tipo de usuario.

Los cuestionarios de evaluación también serán refinados para solicitar más información al usuario sobre el comportamiento del sistema que permita obtener, por ejemplo, información explícita sobre las combinaciones de sistemas de referencia que considera más interesantes.

Por otro lado, el sistema multilingüe necesita todavía ser refinado, el principal problema viene de la traducción de las palabras clave y los términos de realimentación. Se va a utilizar un proceso de desambiguación previo a la traducción para mejorar esta última. En este proceso se van a utilizar como fuentes de información para la desambiguación el contenido de los modelos de usuario donde se encuentran los términos a traducir, ya que se puede suponer que el significado de los términos va a estar en el contexto de los modelos. En todo caso, se va a realizar una evaluación mucho más detallada del sistema multilingüe, con más usuarios y durante más días, obteniendo juicios de los propios usuarios, para poder contrastar las técnicas empleadas en esta propuesta, aunque inicialmente los resultados son prometedores.

Por último, otra mejora interesante del sistema, como fue propuesto en los comentarios de la evaluación cualitativa del último sistema de personalización, será la utilización de resúmenes multidocumento personalizados que permitirán agrupar en un único resumen la información contenida en varias noticias que traten sobre el mismo tema. También la utilización de varios periódicos como fuentes de información permitirá que la información recibida se muestre desde distintos puntos de vista.

BIBLIOGRAFÍA

[Aas97] Aas, K., 1997. "A Survey on Personalised Information Filtering Systems for the World Wide Web". *Report no. 922*, Norwegian Computing Center.

[Acero&Alcojor01] Acero, I. & Alcojor, M., 2001. "Proyecto Pinakes: Servicio personalizado de acceso a la información a través de dispositivos Wireless". *Proyecto fin de carrera.* Escuela Superior de Informática. Universidad Europea de Madrid.

[Acero *et al.* 01] Acero, I., Alcojor, M., Díaz, A., Gómez, J.M., Maña, M., 2001. "Generación automática de resúmenes personalizados". *Procesamiento del Lenguaje Natural*, 27 (2001), pp. 281-290

[Aggarwal *et al.* 99] Aggarwal, C.C., Wolf, J.L., Wu, K., Yu, P.S., 1999. "Horting Hatches an Egg: A New Graph-theoretic Approach to Collaborative Filtering". *Proceedings of the ACM KDD'99 Conference,* pp. 201-212. San Diego, CA.

[Alonso *et al.* 03] Alonso, L., Castellón, I., Climent, S., Fuentes, M., Padró, L., Rodríguez, H., 2003. "Approaches to Text Summarization: Questions and Answers". *Revista Iberoamericana de Inteligencia Artificial,* No. 20, pp. 34-52.

[Alspector *et al.* 98] Alspector, J., Aleksander, K., Karunanithi, N., 1998. "Comparing feature-based and clique-based user models for movie selection", *Proceedings of the Third ACM Conference on Digital Libraries*, pp. 11-18.

[Amato&Straccia99] Amato, G. & Straccia,U., 1999. "User Profile Modeling and Applications to Digital Libraries". Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries (ECDL'99), LNCS 1696, Springer-Verlab, pp. 184-197.

[Ardissono *et al.* 99a] Ardissono, L., Console, L., Torre, I., 1999. "Exploiting user models for personalizing news presentations". *Proc. 2nd Workshop on adaptive systems and user modeling on the World Wide Web*, pp. 13-20, Banf, Canada.

[Ardissono *et al.* 99b] Ardissono, L., Console, L., Torre, I., 1999. "On the application of personalization techniques to news servers on the WWW". *Lecture Notes in Artificial Intelligence 1792*, pp. 261-272.

[Ardissono *et al.* 01] Ardissono, L., Console, L., Torre, I. 2001. "An adaptive system for the personalized access to news". *AI Communications*, vol. 14, N. 3, pp. 129-147.

[Artale *et al.* 97] Artale, A., Magnini, B., Strapparava, C., 1997. "WordNet for Italian and its Use for Lexical Discrimination". *Proceedings of AI*IA97: Advances in Artificial Intelligence.* Springer Verlab.

[Asnicar&Tasso97] Asnicar, F. & Tasso, C., 1997. "ifWeb: a Prototype of User-Model Based Intelligent Agent for Document Filtering and Navigation in the World Wide Web". *Proceedings of the workshop "Adaptive Systems and User Modelling on the World Wide Web"*, 6th International Conference on User Modelling, Chia Laguna, Sardinia, 2-5 June 1997.

[Asnicar *et al.* 97] Asnicar F.A., Di Fant M., Tasso C., 1997. "User Model-Based Information Filtering". *AI*IA 97: Advances in Artificial Intelligence - Proceeding of the 5th Congress of the Italian Association for Artificial Intelligence*, Rome, September 1997, Springer Verlag, Berlin, LNAI 1321, pp. 242-253.

[Attardi *et al.* 99] Attardi, G., Gullí, A., Sebastiani, F., 1999. "Automatic Web Page Categorization by Link and Context Analysis". *Proceedings of THAI-99, 1st European Symposium on Telematics, Hypermedia and Artificial Intelligence*, pp. 105-119.

[Bach *et al.* 96] Bach, J.R., Fuller, C., Gupta, A., Hampapur, A., Horovitz, B., Humphrey, R., Jain, R.C., & Shu, C., 1996. "The VIRAGE Image Search Engine: An Open Framework for Image Management". *Proceedings of the Symposium on Electronic Imagic: Science and Technology-Storage and Retrieval for Image and Video Databases IV, IS&T/SPIE.*

[Baeza-Yates&Ribeiro-Nieto99] Baeza-Yates, R. and Ribeiro-Neto, B., 1999. *Modern Information Retrieval*. ACM Press Books, New York.

[Balabanovic98a] Balabanovic, M., 1998. "Exploring versus Exploiting when Learning User Models for Text Recommendation". *User Modeling and User-Adapted Interaction Journal* 8, pp. 71-102.

[Balabanovic98b] Balabanovic, M., 1998. "An Interface for Learning Multi-topic User Profiles from Implicit Feedback". *AAAI Workshop on Recommender Systems,* Madison, Wisconsin.

[Balabanovic&Shoham97] Balabanovic, M., Shoham, Y., 1997. "Fab: Content-based, collaborative recommendation", *Communications of the ACM,* 40 (3).

[Ballesteros&Croft97] Ballesteros, L. & Croft, W. B., 1997. "Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval". *Research and Development in Information Retrieval*, pp. 84-91.

[Ballesteros&Croft98] Ballesteros, L. & Croft, B.W., 1998. "Statistical methods for cross languages information retrieval". *Cross-Language Information Retrieval*, Kluwer Academir Publishers.

[Barry&Schamber98] Barry, C. L. & Schamber, L., 1998. "User´s criteria for relevance evaluation: a cross-situational comparison". *Information Processing & Management*, 34 (2/3), pp. 219-236.

[Barzilay&Elhadad99] Barzilay, R. & Elhadad, M., 1999. "Text summarizations with lexical chains". *Advances in Automatic Text Summarization*. MIT Press.

[Beaulieu03] Beaulieu, M., 2003. "Approaches to user-based studies in information seeking and retrieval: a Sheffield perspective". *Journal of Information Science*, 29 (4), pp. 239-248.

[Belkin90] Belkin, N. J., 1990. "The cognitive viewpoint in Information Science". *Journal of Information Science*, 16 (1), pp. 11-16.

[Belkin97] Belkin, N. J., 1997. "User Modeling in Information Retrieval". Tutorial disponible en http://www.scils.rutgers.edu/~belkin/um97oh/, *Sixth International Conference on User Modeling, UM97*, Chia Laguna, Sardinia, Italy.

[Belkin00] Belkin, N. J., 2000. "Helping People Find What They Don't Know", *Communications of the ACM,* 43(8), pp. 58-61.

[Belkin01] Belkin, N. J., 2001. "Iterative exploration, design and evaluation of support for query reformulation interactive information retrieval", *Information Processing & Management,* vol. 37, pp. 403-434.

[Benaki *et al.* 97] Benaki, E., Karkaletsis, V., Spyropoulos, C., 1997. "User Modelling in WWW: the UMIE Prototype". *Proceedings of the workshop "Adaptive Systems and User Modelling on the World Wide Web"*, 6th International Conference on User Modelling, Chia Laguna, Sardinia, 2-5 June 1997.

[Berry *et al.* 95] Berry, M.W., Dumais, S.T., O'Brien, G.W., 1995. "Using linear algebra for intelignet information retrieval", *SIAM Review*, Vol. 37, pp. 573-595.

[Bharat *et al.* 98] Bharat, K., Kamba, T., Albers, M., 1998. "Personalized, interactive news on the web". *Multimedia Systems*, (6):349-358.

[Billsus&Pazzani99a] Billsus, D. & Pazzani. M.J., 1999. "A Personal Agent that Talks, Learns and Explains". *Proceedings of the Third International Conference on Autonomous Agents,* pp. 268-275, Seattle, WA.

[Billsus&Pazzani99b] Billsus, D. & Pazzani. M.J., 1999. "A Hybrid User Model for News Story Classification". *Proceedings of the Seventh International Conference on User Modeling, UM99*, 99-108, Banff, Canada.

[Billsus&Pazzani00] Billsus, D. and Pazzani. M.J., 2000. "User Modelinf for Adaptive News Access", *User Modeling and User-Adapted Interaction Journal* 10(2-3), pp. 147-180.

[Borrego99] Borrego, Á., 1999. "La investigación cualitativa y sus aplicaciones en Biblioteconomía y Documentación". *Revista Española de Documentación Científica*, 22 (2), pp. 139-156.

[Boughanem *et al.* 02] Boughanem, M., Chrisment, C., Nassr, N., 2002. "Investigation on Disambiguation in CLIR Aligned Corpus and Bi-directional Translation-Based Strategies". *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, LNCS 2406, Springer. Verlab, pp. 158-168.

[Brandow *et al.* 95] Brandow, R., Mitze, K., Rau, L. F., 1995. "Automatic Condensation of Electronic Publications by Sentence Selection", *Information Processing and Management* 31(5), pp. 675-685.

[Breese *et al.* 98] Breese, J.S., Heckerman, D., Kadie, C., 1998. "Empirical Analysis of Predictive Algorithms for Collaborative Filtering". *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence,* pp. 43-52.

[Buenaga96] Buenaga, M., 1996. Integración de ténicas de Procesamiento de Lenguaje Natural para la Recuperación de Información en bibliotecas de componentes software. *Tesis Doctoral*. Departamento de Informática y Automática, Universidad Complutense de Madrid.

[Buenaga *et al.* 97] Buenaga, M., Gómez, J.M., Díaz B., 1997. "Using WordNet to Complement Training Information in Text Categorization". *Proceedings of the Second International Conference on Recent Advances in Natural Language Processing (RANLP).*

[Buenaga *et al.* 01] Buenaga, M., Maña, M., Díaz, A., Gervás, P., 2001. "A User Model Based on Content Analysis for the Intelligent Personalization of a News Service". *Proceedings of the 8th International Conference on User Modeling (UM'2001)*, LNAI 2109, Springer Verlab, pp. 216-218

[Bueno01] Bueno, D., David, A., 2001. "METIORE: A Personalized Information Retrieval System". *Proceedings of the 8th International Conference on User Modeling (UM'2001)*, LNAI 2109, Springer Verlab, pp. 168-177.

[Bueno02] Bueno, D., 2002. Recomendación personalizada de documentos en sistemas de recuperación de la información basada en objetivos. *Tesis Doctoral*. Departamento de Lenguajes y Ciencias de la Computación. Universidad de Málaga.

[Callan96] Callan, J., 1996. "Document Filtering with Inference Networks". *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 4-11. Zurich, Switzerland, August 1996.

[Carbonell *et al.* 97] Carbonell, J. G., Yang, Y., Frederking, R. E., Brown, R. D., Geng, Y., Lee, D., 1997. "Translingual Information Retrieval: A Comparative Evaluation". *IJCAI (1)*, pp. 708-715.

[Caro-Castro03] Caro-Castro, C., Cediera, L., Travieso, C., 2003. "La investigación sobre recuperación de información desde la perspectiva centrada en el usuario: métodos y variables". *Revista Española de Documentación Científica*, 26 (1), pp. 40-55.

[Chen&Sycara98] Chen, L. & Katia, S., 1998. "WebMate: A Personal Agent for Browsing and Searching". *Proceedings of the Second International Conference on Autonomous Agents*, pp. 132-139, Minneapolis, USA.

[Chesnais *et al.* 95] Chesnais, P., Mucklo, M., Sheena, J., 1995. "The fishwrap personalized news system", *IEEE Second International Workshop on Community Networkins Integrating Multimedia Services to the Home.*

[Chin89] Chin, D. N., 1989, "KNOME: Modeling What the User Knows in UC". *User Models in Dialog Systems*, pp. 74-107.

[Chiu&Webb98] Chiu, B. & Webb, G., 1998. "Using decision trees for agent modeling: improving prediction performance", *User Modeling and User-Adapted Interaction* (8), pp. 131-152.

[Claypool *et al.* 99] Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M., 1999. "Combining content-based and collaborative filters in an online newspaper". *ACMSIGIR Workshop on Recommender Systems*. Berkeley, CA.

[Codina00] Codina, Ll., 2000. "Evaluación de recursos digitales en línea: conceptos, indicadores y métodos". *Revista Española de Documentación Científica*, 23 (1), pp. 9-44.

[Cohen&Singer96] Cohen, W.W. & Singer, Y., 1996. "Context-sensitive learning methods for text categorization". *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, pp. 307-315. ACM Press, New York, US.

[Correia&Boavida02] Correia, N. & Boavida, M., 2002. "Towards and Integrated Personalization Framework: A Taxonomy and Work Proposals". *Proceedings of the AH'2002 Workshop on Personalization Techniques in Electronic Publishing*, pp. 9-18, Málaga, Spain, May 2002.

[Cruz *et al.* 03] Cruz, R.A.P.P. da, García, F.J., Alonso, L., 2003. "Perfiles de usuario: en la senda de la personalización". Informe técnico. DPTOIA-IT-2003-001. Departamento de Informática y Automática. Universidad de Salamanca.

[Dalrymple01] Dalrymple, P. W., 2001. "A quarter century of user-centered study: the impact of Zweizig and Dervin on LIS research". *Library & Information Science Research*, 23 (2), pp. 155-156.

[Davis97] Davis, M., 1997. "New Experiments in Cross-Language Text Retrieval at NMSU's Computing Research Lab". *Proceedings of TREC5,* pp. 447-454. NIST, Gaithesburg, MD.

[Davis98] Davis, M.W., 1998. "On the effective use of large parallel corpora in cross-language text retrieval". *Cross-Language Information Retrieval*, Kluwer Academir Publishers.

[Deerwester *et al.* 90] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. "Indexing by latent semantic analysis". *Journal of the American Society for Information Science, 41,* pp. 391-407.

[Díaz01] Díaz, A., 2001. "Integrating multilingual text classification tasks and user modeling in personalized newspaper services". *Proceedings of the 8th International Conference on User Modeling (UM'2001) (Doctoral Consortium)*, LNAI 2109, Springer Verlab, pp. 268-270.

[Díaz&Gervás00] Díaz, A., Gervás, P., 2000. "Three Information Filtering Applications on the Internet Driven by Linguistic Techniques". *Revue Francaise de Linguistique Appliquee,*V-2(2000), pp. 137-149.

[Díaz&Gervás03] Díaz, A. & Gervás, P., 2003. "Modelado dinámico de usuario en un sistema de personalización de contenidos Web". *Actas de la X Conferencia de la Asociación Española para la Inteligencia Artificial,* pp. 45-54 (Volumen II), San Sebastián, Noviembre 2003.

[Díaz&Gervás04a] Díaz, A. & Gervás, P., 2004. "Dynamic user modeling in a system for personalization of web contents". *Current Topics in Artificial Intelligence, CAEPIA-TTIA 2003*, LNAI 3040, pp. 281-290.

[Díaz&Gervás04b] Díaz, A. & Gervás, P., 2004. "Item Summarization in Personalization of News Delivery Systems". *Text, Speech and Dialogue, TSD 2004*, LNAI 3206, pp. 49-56.

[Díaz&Gervás04c] Díaz, A. & Gervás, P., 2004. "Adaptive User Modeling for Personalization of Web Contents". *Adaptive Hypermedia and Adaptive Web-Based Systems, (Proceedings of AH 2004)*, LNCS 3137, pp. 65-75.

[Díaz *et al.* 00a] Díaz, A., Gervás, P., García A., 2000. "Evaluating a User-Model Based Personalisation Architecture for Digital News Services". *Research and Advanced Technology for Digital Libraries (ECDL00)*, LNCS, Springer Verlag, pp. 259-268

[Díaz *et al.* 00b] Díaz, A., Gervás, P., Gómez, J.M., García, A., Buenaga, M., Chacón, I., San Miguel, B., Murciano, R., Puertas, E., Alcojor, M., Acero, I., 2000. "Proyecto Mercurio: un servicio personalizado de noticias

basado en técnicas de clasificación de texto y modelado de usuario". *Procesamiento del Lenguaje Natural*, 26, pp. 249-250.

[Díaz *et al.* 01a] Díaz, A., Gervás, P., García, A., Chacón, I., 2001. "Sections, Categories and Keywords as Interest Specification Tools for Personalised News Services". *Online Information Review*, 25, no 3 (2001), pp. 149-159.

[Díaz *et al.* 01b] Díaz, A., Maña, M., Buenaga, M, Gómez, J.M., Gervás, P., 2001. "Using linear classifiers in the integration of user modeling and text content analysis in the personalization of a Web-based Spanish News Service". *Workshop on Machine Learning, Information Retrieval and User Modeling*, Julio, 2001, Sonthofen, Alemania

[Díaz *et al.* 01c] Díaz, A., Buenaga, M., Giráldez, I., Gómez, J.M., García, A., Chacón, I., San Miguel, B., Puertas, E., Murciano, R., Alcojor, M., Acero, I., Gervás, P., 2001. "Hermes: Servicios de Personalización Inteligente de Noticias mediante la Integración de Análisis Automático del Contenido Textual y Modelado de Usuario con Capacidades Bilingües". *Procesamiento del Lenguaje Natural*, 27, pp. 299-300

[Díaz *et al.* 02] Díaz, A., Gervás, P., García, A., 2002. "Improving Access to Multilingual Enterprise Information Systems with User Modelling". *Proceedings of the 4th Intenrational Conference on Enterprise Information Systems (ICEIS-2002),* pp. 482-487, Ciudad Real, España, Abril 2002.

[Díaz *et al.* 03] Díaz, A., Gervás, P., García, A., 2003. "Desarrollo de una colección de evaluación para personalización de periódicos digitales". *Actas de las II Jornadas de Tratamiento y Recuperación de la Información (JOTRI 2003),* pp. 10-17, Leganés, Madrid, España, Septiembre 2003.

[Díaz *et al.* 05a] Díaz, A., Gervás, P., García, A., 2005. "System-oriented Evaluation for Multi-tier Personalization of Web Contents". *Proceedings of the Workshop on Intelligent Information Processing*, Las Palmas, España, Febrero 2005 (aceptado).

[Díaz *et al.* 05b] Díaz, A., Gervás, P., García, A., 2005. "Evaluation of a System for Personalized Summarization of Web Contents". *Proceedings of UM2005 User Modeling: Proceedings of the Tenth International Conference*", LNAI, Edinburgh, July 2005 (aceptado).

[Dumais04] Dumais, S., 2004. "Latent semantic analysis". En *ARIST (Annual Review of Information Science Technology)*, volumen 38, pg. 189-230.

[Dumais *et al.* 88] Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwester, S., Harshman, 1998. "Using latent semantic analysis to improve access to textual information". En *Proceedings of CHI'88 Conference on Human Factors in Computing,* pp. 281-285.

[Edmundson69] Edmundson, H.P., 1969. "New Methods in Automatic Abstracting". *Journal of the ACM*, 16(2), pp. 264-285.

[Elliot87] Elliot, C., 1987. "Implementing Web-Based Intelligent Tutors". *Proceedings of the workshop Adaptive Systems and User Modelling on the World Wide Web*, Chia Laguna, Sardinia, June 1997.

[Ellis96] Ellis, D., 1996. *Progress & problems in information retrieval*, London, Library Association.

[Ellis *et al.* 93] Ellis, D., Cox, D, Hall, K., 1993. "A comparison of the information seeking patterns or researchers in the physical and social sciences". *Journal of Documentation*, 49 (4), pp. 356-369.

[Erbach *et al.* 97] Erbach, G., Neumann, G., Uszkoreit, H., 1997. "MULINEX: Multilingual Intexing, Navigation and Editing Extensions for the World-Wide Web". *AAAI Symposium on Cross-Language Text and Speech Retrieval.*

[Fink&Kobsa00] Fink, J. & Kobsa, A., 2000. "A Review and Analysis of Commercial User Modeling Servers for Personalization on the World Wide Web". *User Model and User-Adapted Interaction*, 10(2-3), pp. 209–249.

[Fernández&Moya-Anegón02] Fernández, J. C. & Moya-Anegón, F., 2002. "Perspectivas epistemológicas "humanas" en la Documentación". *Revista Española de Documentación Científica*, 25 (3), pp. 241-253.

[Fidel93] Fidel, R., 1993. "Qualitative methods in information retrieval research". *Library & Information Science Research*, 15 (3), pp. 219-247.

[Fink *et al.* 96] Fink J, Kobsa A, Nill A., 1996. "User-Oriented Adaptivity and Adaptability in the AVANTI project". *Designing for the Web : Empirical Studies*, Microsoft Usability Group, Redmond (WA), 1996

[Frakes&Baeza92] Frakes, W. & Baeza, R., 1992. *Information retrieval: data structures and algorithms*, Prentice Hall, London.

[Gachot *et al.* 98] Gachot, D.A., Lange, E., Yang, J., 1998. "The SYSTRAN NLP browser: An application of machine translation technology in multilingual information retrieval". *Cross-Language Information Retrieval*, Kluwer Academir Publishers.

[García *et al.* 00a] García, A., Chacón, I., Díaz, A., Gervás, P., 2000. "Nuevos sistemas de información: tendencias y evaluación", *Cuadernos de Documentación Multimedia*, nº9, http://www.ucm.es/info/multidoc/multidoc/revista/num9/prensa/jime-chacon.htm

[García *et al.* 00b] García, A., Chacón, I., Díaz, A., Gervás, P., 2000. "Sistemas de información en Internet: estudio de un caso", *Investigación Bibliotecológica*, No. 29 Vol. 14 julio / diciembre de 2000, pp. 114-129

[García *et al.* 00c] García, A., Chacón, I., Díaz, A., Gervás, P., 2000. "Mercurio: Information Professionals in Personalised Digital Information", *Workshop 7: New Professional Fields*, ECN Conference Málaga 2000, "Innovation and change: developing competences for the media and communication professions", Universidad de Málaga, Mayo 2000.

[García *et al.* 02] García, A., Díaz, A., Gervás, P., 2002. "Knowledge Organization in a Multilingual System for the Personalization of Digital News Services: a way of integration of the knowlege". *Advances in Knowledge Organization,* Vol. 8 (2002), Ergon Verlab, pp. 386-392.

[Gates *et al.* 98] Gates, K.F., Lawhead, P.B., Wilkins, D.E., 1998. "Towards an adaptive WWW: a case study in customized hypemedia". *The New Review of Hypermedia and Multimedia,* 4.

[Gervás *et al.* 99] Gervás, P., San Miguel, B., Díaz, A., García, A., 1999. "Mercurio: un servidor personalizado de noticias basado en modelos de usuario obtenidos a través de la WWW"*, III Congreso de Investigadores Audiovisuales (Los medios del tercer milenio)*, Noviembre 1999,

Facultad de Ciencias de la Información, Universidad Complutense de Madrid.

[Gey&Oard01] Gey, F. C. & Oard, D. W., 2001. "The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic using Englis, French or Arabic Queries". *Proceedings of TREC10*. NIST, Gaithesburg, MD.

[Gilarranz97] Gilarranz, J., Gonzalo, J., Verdejo, M.F., 1997. "Language-Independent Text Retrieval using the EuroWordNet Multilingual Semantic Database". *Proceedings of the Second Workshop on Multilinguality in the Software Industry: the AI contributioin (IJCAI97).*

[Giráldez *et al.* 02] Giráldez, I., Puertas, E., Gómez, J.M., Murciano, R., Chacón, I., 2002. "Hermes: Intelligent Multilingual News Filtering based on Language Engineering for Advanced User Profiling", *Multilingual Information Access and Natural Language Processing Workshop Proceedings*, *VIII Iberoamerican Conference on Artificial Intelligence (IBERAMIA)*, pp. 81-88.

[Good *et al.* 99] Good, N., Schafer, J.B., Konstan, J.A., Borchers, A., Sarwar, B.M., Herlocker, J.L., Riedl, J., 1999. "Combining Collaborative Filtering with Personal Agents for Better Recommendations". *Proceedings of the Sixteenth National Conference on Artificial Intelligence AAAI/IAAI*, pp. 439-446.

[Goldberg *et al.* 92] Goldberg, D., Nichols, D., Oki, B. M., Terry, D., 1992. "Using collaborative filtering to weave an information tapestry". *Communications of the ACM*, 35(12), pp. 61-70.

[Golub&Loan96] Golub, G.H. & Van Loan, C.F., 1996. *Matrix Computations 3rd. ed.* The Johns Hopkins University Press.

[Gómez *et al.* 01] Gómez, J.M., Murciano, R., Díaz, A., Buenaga, M., Puertas, E., 2001. "Categorizing photographs for user-adapted searching in a news agency e-commerce application", *1st International Workshop on New Developments in Digital Libraries (NDDL-2001)*, Julio 2001, Setúbal, Portugal.

[Gómez&Buenaga97] Gómez, J.M. & Buenaga, M., 1997. "Integrating a Lexical Database and a Training Collection for Text Categorisation". *ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP*, Madrid, España.

[Gonzalo *et al.* 98] Gonzalo, J., Verdejo, F., Chugur, I., Cigarran, J., 1998. "Indexing with WordNet synsets can improve Text Retrieval". *Proceedings of the COLING/ACL'98 Workshop on Usage of WordNet for NLP*, Montreal, 1998.

[Gonzalo *et al.* 99] Gonzalo, J., Verdejo, F., Chugur, I., 1999. "Using EuroWordNet in a concept-based approach to cross-language text retrieval". *Applied Atificial Intelligence*, 13(7), pp. 647-678.

[González *et al.* 02] González, J.C., Villena, J., Bueno, F., García Serrano, A.M., Martínez, P., 2002. "OMNIPAPER: Acceso Inteligente a Diarios de la Unión Europea". *Procesamiento de Lenguaje Natural* 29, pp. 289-290.

[Grefenstette98a] Grefenstette, G. (ed.), 1998. *Cross-Language Information Retrieval*, Kluwer Academir Publishers.

[Grefenstette98b] Grefenstette, G., 1998. "The Problem of Cross-Language Information Retrieval". *Cross-Language Information Retrieval*, Kluwer Academir Publishers.

[Grefenstette99] Grefenstette, G., 1999. "The WWW as a Resource for Example-Based MT Tasks". *ASLIB'99, Translating and the Computer, 21,* London, UK, Nov 1999.

[Hahn&Reimer99] Hahn, U. & Reimer, U., 1999. "Knowledge-Based Text Summarization: Salience and Generalization Operators for Knowledge-Based Abstraction". *Advances in Automatic Text Summarization*, Cambridge, Massachusetss, The MIT Press.

[Hahn&Mani00] Hahn, U. & Mani, I., 2000. "The Challenges of Automatic Summarization". *Computer*, 33(11), pp.29-36.

[Hanani *et al.* 01] Hanani, U., Shapira, B., Shoval, P., 2001. "Information Filtering: Overview of Issues, Research and Systems", *User Modeling and User-Adapted Interaction,* 11(3), pp. 203-259.

[Harman93] Harman, D.K., 1993. "The first Text REtrieval Conference (TREC-1)", *Information Processing and Management*, 29(4), pp. 411-414.

[Harris86] Harris, M., 1986. "The dialectic of defeat: antinomies in research in library and information science". *Library Trends*, 34 (3), pp. 515-531.

[Hearst97] Hearst, M., 1997. "TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages". *Computational Linguistics*, Vol. 23 No. 1, pp. 33-64.

[Hearst99] Hearst, M., 1999. "User Interfaces and Visualization". *Modern Information Retrieval*. ACM Press Books, New York.

[Hyldegaard&Seiden04] Hyldegaard, J. & Seiden, P., 2004. "My e-journal – exploring the usefulness of personalized access to scholarly articles and services". *Information Research*, 9 (3).

[Hull93] Hull, D., 1993. "Using statistical testing in the evaluation of retrieval experiments". *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 329-338.

[Hull94] Hull, D.A., 1994. "Improving text retrieval for the routing problem using latent semantic indexing". *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, pp. 282-289, Springer Verlag, Heidelberg, DE.

[Hull98] Hull, D., 1998. "A weighted boolean model for cross-language text retrieval". *Cross-Language Information Retrieval*, Kluwer Academir Publishers.

[Ingwersen96] Ingwersen, P., 1996. "Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory". *Journal of Documentation*, 52 (1), pp. 3-50.

[Jennings&Higuchi92] Jennings, A. & Higuchi, H., 1992. "A personal news service based on a user model neural network". *IEICE Transactions on Information and Systems*, March 1992.

[Johnson *et al.* 03] Johnson, F. C., Griffiths, J. R., Hartley, R. J., 2003. "Task dimensions of user evaluations of information retrieval systems". *Information Research*, 8 (4).

[Jones *et al.* 00] Jones, G.J.F., Quested, D.J., Thomson, K.E., 2000. "Personalised delivey of news articles from multiple sources". Research and Advanced Technology for Digital Libraries (ECDL00), LNCS, Springer Verlag, pp. 340-343.

[Kamba *et al.* 95] Kamba, T., Bharat, K., Albers, M., 1995. "The krakatoa chronicle: An interactive, personalized newspaper on the web". *Proceedings of the Fourth International World-Wide Web Conference*, pp. 159-170. Boston, MA, December 1995.

[Kilander *et al.* 97] Kilander, F., Fåhraeus, E., Palme, J., 1997. "PEFNA The private filtering news agent". *Technical report 97-004*, Department of Computer and Systems Sciences, Stockholm University.

[Kobsa *et al.* 01] Kobsa, A., Koenemann, J., Wolfgang, P., 2001. "Personalised hypermedia presentation techniques for improving online customer relationships". *The Knowledge Engineering Review*, Vol. 16:2, Cambridge University Press, pp. 111–155.

[Konstan *et al.* 97] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., Riedl, J., 1997. "GroupLens: Applying Collaborative Filtering to Usenet News". *Communications of the ACM,* 40(3), pp. 77-87.

[Kron *et al.* 96] Kroon, H.C.M. de, Mitchell, T.M., Kerckhoffs, E.J.H., 1996. "Improving learning accuracy in information filtering". *Procceding of the International Conference on Machine Learning - Workshop on Machine Learning Meets HCI* (ICML-96), 1996.

[Kupiec *et al.* 95] Kupiec, J., Pedersen, J.O., Chen, F., 1995. "A Trainable Document Summarizer", *Research and Development in Information Retrieval*, pp. 68-73.

[Labrou&Finin00] Labrou, Y. & Finin, T., 2000. "Yahoo! As an Ontology: Using Yahoo! Categories to Describe Documents". *Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM-99)*, pp. 180-187, ACM Press, November 2-6 2000.

[Lewis92] Lewis, David D., 1992. *Representation and Learning in Information Retrieval.* PhD Thesis, Technical Report UM-CS-1991-093, Department of Computer and Information Science, University of Massachussetts.

[Lewis98] Lewis, D.D., 1998. "Naive (Bayes) at forty: The independence assumption in information retrieval". *Proceedings of ECML-98, 10th European Conference on Machine Learning*, Lecture Notes in Computer Science, Number 1398, pp. 4-15, Springer Verlag, Heidelberg, DE.

[Lewis&Ringuette94] Lewis, D.D. & Ringuette, M., 1994. "A comparison of two learning algorithms for text categorization". *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 81-93.

[Lewis *et al.* 96] Lewis, D.D., Schapire, R.E., Callan, J.P., Papka, R., 1996. "Training algorithms for linear text classifiers". *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, pp. 298-306, ACM Press, New York, US.

[Lieberman *et al.* 01] Lieberman, H., Fry, C., Weitzman, L., 2001. "Exploring the Web with Reconnaissance Agents".*Communications of the ACM*, Vol 44., No. 8, pp. 69-75.

[Lieberman95] Lieberman, H., 1995. "Letizia: An agent that assists Web browsing". *Proceeding of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI95)*, pp. 924-929.

[LópezOstenero *et al.* 04] López-Ostenero, F., Gonzalo, J, Verdejo, F., 2004. "Búsqueda de información multilingüe: estado del arte". *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial,* Vol. VIII, No 22, pp. 11-35.

[Luhn58] Luhn, H.P., 1958. "The Automatic Creation of Literature Abstracts". *IBM Journal of Research and Development*, 2(2), pp. 159-165.

[Luque *et al.* 99] Luque, V., Fernández, C., Delgado, C., 1999. "Personalizing your electronic newspaper". *Proceedings of the 4th Euromedia Conference (WEBTEC)*, pp. 26-28.

[Magnini&Strapparava00] Magnini, B. & Strapparava, C., 2000. "Experiments in word domain disambiguation for parallel texts". *Proceedings of the SIGLEX Workshop on Word Senses and Multilinguality*, Hong Konk, Octeber 2000.

[Magnini&Strapparava01] Magnini, B. & Strapparava, C., 2001. "Improving User Modeling with Content-Based Techniques". *Proceedings of the 8th Internantional Conference on User Modeling (UM01).* Springer Verlab, pp. 74-83.

[Mani01] Mani, I., 2001. *Automatic Summarization.* John Benjamins Publishing Company, Amsterdam/Philadelphia.

[Mani&Maybury99] Mani, I. & Maybury, M. (eds.), 1999. *Advances in Automatic Text Summarization.* Cambridge, Massachusetss: The MIT Press.

[Maña03] Maña, M.J., 2003. Generación automática de resúmenes de texto para el acceso a la información. *Tesis doctoral*. Departamento de Informática. Universidad de Vigo. Septiembre 2003.

[Maña *et al.* 99] Maña, M.J., Buenaga, M., Gómez, J.M., 1999. "Using and Evaluating User Directed Summaries to Improve Information Access". *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries (ECDL'99)*, Lecture Notes in Computer Science, Vol. 1696, pp. 198-214, Springer-Verlag.

[Maña *et al.* 00] Maña, M.J., Ureña, L.A., Buenaga, M., 2000. "Tareas de análisis del contenido textual para la recuperación de información con realimentación". *Procesamiento del Lenguaje Natural, 24.* Septiembre 2000.

[Marcu00] Marcu, D., 2000. *The Theory and Practice of Discourse Parsing and Summarization.* The MIT Press.

[Martínez *et al.* 01] Martínez, F., Ureña, A., García, M., 2001. "WWW como Fuente de Recursos Lingüísticos para su Uso en PLN". *Procesamiento del Lenguaje Natural, 27.* Septiembre 2001.

[Mateo *et al.* 03] Mateo, P.L., González, J.C., Villena, J., Martínez, L.L., 2003. "Un sistema para resumen automático de textos en castellano". *Procesamiento de Lenguaje Natural*, núm. 31, págs. 29-36

[Maybury&Mani01] Maybury, M., Mani, I., 2001. *Automatic Summarization.* ACL/EACL'01 Tutorial.

[Miller95] Miller, G.A., 1995. "WordNet: A Lexical Database for English". *Communications of the ACM* 38(11):39-41.

[Mitchell97] Mitchell, T., 1997. *Machine Learning*. McGraw-Hill, New York.

[Mizarro97] Mizzaro, S., 1997. "Relevance: The Whole History". *Journal of the American Society for Information Science,* 48(9), pp. 810-832.

[Mizarro01] Mizzaro, S., 2001. "A New Measure Of Retrieval Effectiveness (or: What's Wrong With Precision And Recall)". In: T. Ojala (ed.): *International Workshop on Information Retrieval (IR'2001)*, Infotech Oulu, pp. 43-52.

[Mizarro&Tasso02] Mizarro, S. & Tasso, C., 2002. "Ephemeral and Persistent Personalization in Adaptive Information Access to Scholarly Publications on the Web". *Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems*, Málaga, España, Mayo 2002.

[Mladenic98a] Mladenic, D., 1998. "Tuning Yahoo into an Automatic Web-Page Classifier". *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI98),* pp. 473-474.

[Mladenic98b] Mladenic, D., 1998. "Feature subset selection in text-learning". *Proceedings of the 10th European Conference on Machine Learning (ECML98)*, pp. 95-100.

[Moreiro02] Moreiro, J. A., 2002. "Criterios e indicadores para evaluar la calidad del análisis documental de contenido". *Ciencia da informaçao*, 31 (1), pp. 53-60.

[Morris *et al.* 92] Morris, J., Kasper, G., Adams, D., 1992. "The Effects and Limitations of Automated Text Condensing on Reading Comprehension Performance". *Information Systems Research*, 3(1), pp. 17-35.

[Morris&Hirst91] Morris, J. & Hirst, G., 1991. "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text". *Computational Linguistics*, Vol. 17, No. 1, pp. 21-48.

[Moukas96] Moukas, A., 1996. "Amalthaea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem". *Proceedings of the Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology*. London, UK, 1996.

[Myaeng&Jang99] Myaeng, S. & Jang D., 1999. "Development and Evaluation of Statistically-Based Document Summarization System". *Advances in Automatic Text Summarization*, MIT Press, pp. 61-70.

[Nakao00] Nakao, Y, 2000. "An Algorithm for One-Page Summarization of a Long Text Based on Thematic Hierarchy Detection". *Proceedings of the 38th Meeting of hte Association for Computational Linguistics,* pp. 302-309.

[Nakamura *et al.* 95] Nakamura, A., Mamizuka, H., Toba, H., Abe, N., 1995. "Learning personal preference functions using boolean-variable real-valued multivariate polynomials". *52nd National Convention of the Information Processing Society of Japan.*

[Nakashima&Nakamura97] Nakashima, T. & Nakamura, R., 1997. "Information Filtering for the Newspaper". *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, August 1997. Victoria, B.C., Canada.

[Nanba&Okumura00] Nanba, H. & Okumura, M., 2000. "Producing More Readable Extracts by Revising Them". *Proceedings of the 18th International Conference on Computational Linguistics,* pp. 1071-1075.

[Niblack *et al.* 93] Niblack, W., Barner, R., Equitz, W., Flickner, M., Glasman, E., Petkovic, D., Yanker, P., Faloutsos, C. and Taubin, G., 1993. "The QBIC Project: Quering Images by Content Using Color, Texture, and Shape". *Proceedings of the Symposium on Electronic Imagic: Science and Technology-Storage and Retrieval for Image and Video Databases*.

[Nie99] Nie, J. Y., 1999. "TREC-7 CLIR using a Probabilistic Translation Mode". *Proceedings of TREC7*, pp. 547-554. NIST, Gaithesburg, MD.

[Oard98] Oard, D. W., 1998. "A comparative study of query and document translation for cross-language information retrieval". *Proceedings of the Third Conference of the Association for Machine Translation in the Americas*.

[Oard01] Oard, D. W., 2001. "Evaluating Interactive Cross-Language Information Retrieval: Document Selection". *Cross- Language Information Retrieval and Evaluation, Workshop, CLEF 2000*, LNCS 2069, Spriger Verlab, pp. 57-71.

[Oard&Dorr98] Oard, D. & Dorr, B.J., 1998. "Evaluating cross-language text filtering effectiveness". *Cross-Language Information Retrieval*, Kluwer Academir Publishers.

[Ørom00] Ørom, A., 2000. "Information Science, historical changes and social aspects: a nordic outlook". *Journal of Documentation*, 56 (1), pp. 12-26.

[Pastor&Asensi99] Pastor, J. A. & Asensi, V., 1999. "Un modelo para la evaluación de interfaces en sistemas de recuperación de información", *IV congreso Isko-España (Eoconsid'99), La Representación y la Organización del Conocimiento en sus distintas perspectivas: su influencia en la Recuperación de la Información,* Isko-España. Universidad de Granada, pp. 401-409.

[Paice90] Paice, C.D., 1990. "Constructing Literature Abstracts by Computer: Techniques and Prospects". *Information Processing and Management*, 26(1), pp. 171-186.

[Pentland *et al.* 94] Pentland, A., Picard, R.W., Sclaroff, S., 1994. "Photobook: Tools for Content-Based Manipulation of Image Databases". *Proceedings of the Symposium on Electronic Imagic: Science and Technology-Storage and Retrieval for Image and Video Databases.*

[Perkowitz&Etzioni00] Perkowitz, M. & Etzioni, O. "Towards adaptive Web sites: Conceptual framework and case study", *Artificial Intelligence* 118, No. 1-2, pp. 245-275.

[Peters01] Peters, C., 2001. "Results of the CLEF 2001 Cross-Language System Evaluation Campaign", *Working Notes for the CLEF 2001 Workshop,* Darmstadt, Germany, September 2001.

[Picchi&Peters98] Picchi, E. & Peters, C., 1998. "Cross-language information retrieval: a system for comparable corpus querying". *Cross-Language Information Retrieval*, Kluwer Academir Publishers.

[Pirkola98] Pirkola, A., 1998. "The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval". *Proceedings of SIGIR'98*, pp. 55-63.

[Porter80] Porter M.F., 1980. "An algorithm for suffix stripping". *Program* 14(3), pp. 130-137.

[Pretschner&Gauch99] Pretschner, A. & Gauch, S., 1999. "Personalization on the Web". *Technical Report ITTC-FY2000-TR-13591-01*. Information and Telecommunication Technology Center. Department of Electrical Engineering and Computer Science. The University of Kansas.

[Quillian68] Quillian, M., 1968. "Semantic memory". In M. Minsky, editor, *Semantic Information Processing*, MIT Press, Cambridge, MA, pp. 227-270.

[Resnick *et al.* 94] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J., 1994. "GroupLens: An open architecture for collaborative filtering of netnews". *Proceedings of the Conference on Computer-Supported Cooperative Work*, pp. 175-186.

[Rich79] Rich, E., 1979. "User modeling via stereotypes". *Cognitive Science* 3, pg. 329-354.

[Rich83] Rich, E., 1983. "Users are individuals: individualising user models". *International Journal of Man-Machine Studies*, 18, pp. 199-214.

[Robertson&Sparck Jones76] Robertson, S.E. & Sparck Jones, K., 1976. "Relevance Weighting of Search Terms". *Journal of the American Society for Information Sciences,* 27(3), pp. 129-146.

[Rocchio71] Rocchio, J.J. Jr., 1971. "Relevance feedback in information retrieval", *The SMART Retrieval System: Experiments in Automatic Document Processing,* Prentice-Hall.

[Röpnack *et al.* 98] Röpnack, A., Schindler, M., Schwan, T., 1998. "Concepts of Enterprise Knowledge Medium". *Proceedings of the 2nd International Conference on Practical Aspects of Knowledge Management (PAKM'98)*, Basel, Switzerland.

[Sakagami&Kamba97] Sakagami, H. & Kamba, T., 1997. "Learning personal preferences on online newspaper articles from user behaviours". *Proceedings of the Sixth International World Wide Web Conference (WWW6),* Santa Clara, CA, pp. 291-300.

[Salampasis&Diamantaras02] Salampasis, M. & Diamantaras, K. I., 2002. "Experimental User-Centered Evaluation of an Open Hypermedia System of an Open Hypermedia System and Web Information Seeking Environments". *Journal of Digital Information*, vol. 2 (4).

[Salton71] Salton, G., 1971. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall.

[Salton89] Salton, G., 1989. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley Publishing, Reading, Massachusets, US, 1989.

[Salton&Buckley88] Salton, G. & Buckley, C., 1988. "Term-weighting approaches in automatic text retrieval". *Information Processing and Management,* 24(5): 513-523.

[Salton&Buckley90] Salton, G. & Buckley, C., 1990. "Improving Retrieval Performance by Relevance Feedback". *Journal of the American Society for Information Science* 41(4), pp. 288-297

[Salton&McGill83] Salton G. & McGill, M.J., 1983. *Introduction to modern information retrieval*. McGraw-Hill, New York.

[Salton *et al.* 76] Salton, G., Wong, A. & Wu, C.T., 1976. "Automatic indexing using term discrimination and term precision measurements". *Information Processing and Management,* 12.

[Salton *et al.* 94] Salton, G., Allan, J., Buckley, C., Singhal, A., 1994. "Automatic Analysis, Theme Generation and Summarization of Machine-Readable Texts". *Science* (264), pp. 1421-1426.

[Saracevic96] Saracevic, T., 1996. "Interactive models in information retrieval (IR): A review and proposal". *Proceddings of the 59th Annual Meeting of the American Society for Information Science*, 33, pp. 3-9.

[Sarwar *et al.* 01] Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl J., 2001. "Item-based Collaborative Filtering Recommendation Algorithms". *Proceedings of the 10 International World Wide Web Conference*. Hong Kong.

[Schwartz98] Schwartz, C., 1998. "Web Search Engines", *JASIS*, vol. 49, n. 11, pp. 973-982.

[Schütze *et al.* 95] Schütze, H., Hull, D.A., Pedersen, J.O., 1995. "A comparison of classifiers and document representations for the routing problem". *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*, pp. 229-237, ACM Press, New York, US.

[Sebastiani99] Sebastiani, F., 1999. "A tutorial on automated text categorisation". *Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence*, pp. 7-35.

[Sebastiani02] Sebastiani, 2002. "Machine learning in automated text categorization". ACM Computing Surveys (CSUR), 34(1), pp. 1-47, ACM Press, 2002.

[Skorochod'ko72] Skorochod'ko, E., 1972. "Adaptive method of automatic abstracting and indexing". *Information Processing 71: Proceedings of the IFIP Congress 71*, pp. 1179-1182. North-Holland Publishing Company.

[Shepherd *et al.* 02] Shepherd, M., Watters, C. and Marath, A.T., 2002. "Adaptive User Modeling for Filtering Electronic News". *35th Hawaii International Conference on System Sciences*, Hawaii, 7-10 January 2002. CD-ROM publication.

[Sheridan&Ballerini96] Sheridan, P. & Ballerini, J.P., 1996. "Experiments in multilingual information retrieval using the SPIDER system". *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp 58-65.

[Sheridan *et al.* 97] Sheridan, P., Braschler, M., Schäuble, P., 1997. "Cross-language information retrieval in a multi-lingual legal domain". *Proceedings of ECDL-97, 1st European Conference on Research and Advanced Technology for Digital Libraries*, pp. 253-268, Pisa, IT.

[Sheth&Maes93] Sheth, B. & Maes, P., 1993. "Evolving agents for personalized information filtering". *Proceedings of the Ninth Conference on Artificial Intelligence for Applications*. IEEE Computer Society Press.

[Signore *et al.* 97] Signore, O., Bartoli, R., Fresta, G., 1997. "Tailoring Web Pages to Users' Needs". *Proceedings of the workshop Adaptive Systems and User Modeling on the World Wide Web, Sixth International Conference on User Modeling,* Chia Laguna, Sardinia, 2-5 June 1997.

[Slype91] Slype, G. Van., 1991. *Los lenguajes documentales de indización: concepción, construcción y utilización en los sistemas documentales*, Madrid. Salamanca, Fundación Germán Sánchez Ruipérez; Pirámide.

[Smith&Chang97] Smith, J.R. & Chang S.-F., 1997. "Visually Searching the Web for Content". *IEEE Multimedia*, 4(3), pp. 12-20.

[Sorensen&Elligott95] Sorensen, H. & Elligott, M.Nc., 1995. "PSUN: A Profiling System for Usenet News". CIKM*95 Intelligent Information Agents Workshop*, Baltimore, December 1995.

[Sparck Jones *et al.* 98] Sparck Jones, K., Walker, S., Robertson., S., 1998. "A probabilistic model of information retrieval: development and status". *Technical Report TR-446*, Cambridge University Computer Laboratory, 1998.

[Sparck Jones99] Sparck Jones, K., 1999. "Automatic Summarizing: Factors and Directions". *Advances in Automatic Text Summarization,* pp. 1-13. The MIT Press.

[Sperer&Oard00] Sperer, R. & Oard, D. W., 2000. "Structured Translation for Cross-Language Information Retrieval". *Proceedings of SIGIR'2000*, pp. 120-127.

[Spink97] Spink, A., 1997. "Study of interactive feedback during mediated information retrieval". *Journal of the American Society for Information Science*, 48 (5), pp. 382-394.

[Spink02] Spink, A., 2002. "A user-centered approach to evaluating human interaction with Web search engines: an exploratory study". *Information Processing & Management*, 38(3), pp. 401-426.

[Spink *et al.* 98] Spink, A., Howard, G., Bateman, J., 1998. "From highly relevant to not relevant: examining different regions of relevance". *Information Processing & Management*, 34 (5), pp. 599-621.

[Spink *et al.* 02] Spink, A., Jansen, B., Wolfram, D., Saracevic, T., 2002. "From E-Sex to E-Commerce: Web Search Changes". *Computer* 35(3), pp.107-109.

[Su92] Su, L. T., 1992. "Evaluation measures for interactive information retrieval". *Information Processing and management*, 28 (4), pp. 503-516.

[Su98] Su, L. T., 1998. "Value of search results as a whole as the best single measure of information retrieval performance". *Information Processing and Management*, 34 (5), pp. 57-579.

[Terveen et al .97] Terveen, L., Hill, W., Amento, B., McDonald, D., Creter, J., 1997. "PHOAKS: a system for sharing recommendations," *Communications of the ACM*, vol. 40, pp. 59-62.

[Teufel&Moens97] Teufel, S. & Moens, M., 1997. "Sentence extraction as a classification task". *Proceedings of ACL/EACL Workshop on Intelligent Scalable Text Summarization*, pp. 58-65, Madrid, España.

[Theng *et al.* 99] Theng, Y.L., Duncker, E., Mohd-Nasir, N., Buchanan, G., Thimbleby, H., 1999. "Design Guidelines and User-Centred Digital Libraries". *Third European Conference on Digital Libraries, ECDL'99*, Paris, France, September 1999.

[Tombros&Sanderson98] Tombros, A. & Sanderson, M., 1998. "Advantages of query-biased Summaries in IR". *Proceedings of the 21st ACM SIGIR Conference*, pp. 2-10.

[Tsybenko&Bikov97] Tsybenko, Y. & Bikov, V., 1997. "Feedback and Adaptive Interaction for WWW-Based Courses". *Proceedings of the workshop Adaptive Systems and User Modelling on the World Wide Web*, Chia Laguna, Sardinia, June 1997.

[Turtle&Croft91] Turtle, H.R. & Croft, W.B., 1991. "Evaluation of an Inference Network-Base Retrieval Model". *ACM Transactions on Information Systems,* vol. 9, num. 3, pp. 188-222.

[Van Dijk77] Van Dijk, T.A., 1977. "Semantic Macro-Structures and Knowledge Frames in Discourse Comprehension". *Cognitive Processes in Comprehension.* Lawrence Erlbaum, Hillsdale, N.J., pp. 3-32.

[Van Rijsbergen77] Van Rijsbergen, C., 1977. "A theoretical basis for the use of co-occurrence data in information retrieval". *Journal of Documentation*, 12 (2).

[Van Setten01] Van Setten, M., 2001. "Personalised Information Systems". *Rep. TI/RS/2001/036.* Enschede, The Netherlands: Telematic Instituut.

[Vargas-Quesada *et al.* 02] Vargas-Quesada, B, Moya, F. de, Olvera, M. D., 2002. "Enfoques en torno al modelo cognitivo para la recuperación de información: análisis crítico". *Ciencia da Informaçao*, 31 (2), pp. 107-119.

[Vassileva97] Vassileva, J., 1997. "Dynamic Courseware Generation on the WWW". *Proceedings of the workshop Adaptive Systems and User Modelling on the World Wide Web*, Chia Laguna, Sardinia, June 1997.

[Veltman98] Veltman, G., 1998. "A multi-agent system for generating a personalized newspaper digest". *AAAI/ICML-98 Workshop on Learning for Text Categorization, Technical Report WS-98-05,* AAAI Press, pp. 99-102.

[Verdejo00] Verdejo, M.F., Gonzalo, J., Peñas, A., López, F, Fernández, D., 2000. "Evaluating wordnets in Cross-Language Text Retreival: the ITEM multilingual search engine". *Proceedings of the Second Language Resources and Evaluation Conference (LREC00), Athens.*

[Vossen96] Vossen, P., 1996. "Eurowordnet: building multilingual wordnet database with semantic relations between words". *Technical report*, EC-funded project LE # 4003.

[Vossen97] Vossen, P., 1997. "EuroWordNet: a multilingual database for information retrieval". *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, March, Zurich.

[Webb *et al.* 01] Webb, G.I., Pazzani, M.J., Billsus, D., 2001. "Machine Learning for User Modeling". *User Modeling and User-Adapted Interaction Journal* 11, pp. 19-29.

[Widyantoro *et al.* 01] Widyantoro, D.H., Ioerger, T.R., Yen, J., 2001. "Learning User Interest Dynamics with a Three-Descriptor Representation". *Journal of the American Society for Information Science and Technology* 52(3), pp. 212-225.

[Xu&Croft96] Xu, J. & Croft, W.B., 1996. "Query Expansion Using Local and Global Document Analysis". *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 4-11. Zurich, Switzerland, August 1996.

[Yan&Garcia-Molina95] Yan, Tak W. & Garcia-Molina, H., 1995. "SIFT – A Tool for Wide-Area Information Dissemination". *Proceedings of the USENIX Technical Conference*, pp. 177-186.

[Yang99] Yang, Y., 1999. "An evaluation of statistical approaches to text categorization". *Information Retrieval*, Vol. 1, Number 1-2, pp. 69-90.

[Yang&Chung04] Yang, C. C. & Chung, A., 2004. "Intelligent infomediary for web financial information". *Decision Support Systems*, 38 (1), pp. 65-80.

[Yang&Pedersen97] Yang, Y. & Pedersen, J.O., 1997. "A comparative study on feature selection in text categorization". *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pp. 412-420, Morgan Kaufmann Publishers, San Francisco, US.

[Yang *et al.* 98] Yang, Y., Carbonell, J. G., Brown, R. D., Frederking, R. E., 1998. "Translingual Information Retrieval: Learning from Bilingual Corpora". *Artificial Intelligence*, 103(1-2), pp.323-345.

Apéndice I
CUESTIONARIOS DE EVALUACIÓN

En este apéndice se van a presentar los distintos cuestionarios de evaluación utilizados a lo largo de la tesis. El primero de ellos se utilizó para realizar la evaluación cualitativa del primer experimento preliminar (apartado I.1) y el segundo se utilizó para evaluar el sistema multilingüe (apartado I.2). Estos dos cuestionarios fueron rellenados por los usuarios al terminar el proceso de evaluación correspondiente. Los dos últimos cuestionarios corresponden a los utilizados en la evaluación cualitativa del sistema de personalización 2.0 (apartado I.3). El primero de ellos corresponde a la evaluación inicial y el segundo a la evaluación final.

I.1. Cuestionario de evaluación utilizado en el primer experimento preliminar

Este cuestionario contiene información sobre las opiniones de los usuarios y sobre los datos recogidos por los usuarios de la evaluación exhaustiva del sistema. Es un documento Word con el siguiente aspecto:

Analista:

Navegador utilizado:

Fecha del análisis (días en los que ha usado el servicio):

Días que ha realizado la comprobación exhaustiva:

Incidencias:

Debe cumplimentar este cuestionario dejando sólo la opción elegida (y borrando las opciones que no queremos). Un ejemplo:

¿El sistema es fácil de instalar, usar y mantener?

BUENO

A) CUESTIONARIO

1) Evaluación de la interfaz (análisis de la calidad)

a) Arquitectura de acceso:

- ¿El sistema es fácil de instalar, usar y mantener?

ÓPTIMO/BUENO/REGULAR/MALO

b) Interfaz General:

- ¿Cuál es el grado de satisfacción con los componentes gráficos: iconos o barras de herramientas que permitan el acceso a todas las funciones?

ÓPTIMO/BUENO/REGULAR/MALO

c) Adaptación al usuario:

- ¿Es amigable para el usuario darse de alta y de baja de suscripción?

ÓPTIMO/BUENO/REGULAR/MALO

d) Gestión de contenidos:

- ¿Presenta el sistema la lista de las diferentes categorías y/o secciones en la suscripción? SI/NO

- ¿Permite la utilización de dos categorías y/o secciones al mismo tiempo? SI/NO

- ¿Se puede pasar de una categoría y/o sección a otra? SI/NO

- ¿Permite elegir los días de la semana en que se quiere recibir la información? SI/NO

- ¿El sistema de enlaces es bueno? (colores, longitud, título, etc.)

ÓPTIMO/BUENO/REGULAR/MALO

e) Esquemas de búsqueda:

- ¿Existe un sistema de búsqueda que le permita obtener noticias en el periódico?: SI/NO

- ¿Permite la interfaz que el usuario presente o modifique sus criterios de búsqueda?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿Ofrece un historial de búsquedas y de documentos relevantes?

- ÓPTIMO/BUENO/REGULAR/MALO

- ¿El sistema ofrece el número de documentos incluidos en la recuperación y ofrece posibilidades de limitación?

ÓPTIMO/BUENO/REGULAR/MALO

f) Esquemas de recuperación y consulta:

- ¿Ofrece la posibilidad de ordenar los resultados por diferentes criterios (relevancia, por campos, etc.)? **En caso negativo ponga NO**

ÓPTIMO/BUENO/REGULAR/MALO

- ¿Se resaltan los elementos ya consultados con anterioridad? SÍ/NO

- ¿Permite acceder a los documentos desde cualquier sitio que se mencione? SÍ/NO

- ¿Cuál es el esfuerzo y el tiempo requerido para la recuperación?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿Permite el sistema observar la frecuencia de términos en el documento relacionados con la categoría? SÍ/NO

- ¿Permite el sistema el acceso al documento completo? SÍ/NO

- ¿El sistema permite la localización de la categoría en el resumen y en el documento? SÍ/NO

g) Sistemas de ayuda al usuario:

- Valore la introducción al sistema. ÓPTIMO/BUENO/REGULAR/MALO

- ¿Se proporciona información sobre la actividad realizada cuando la acción requiere un tiempo considerable? SÍ/NO

- ¿Proporciona mensajes de error? SÍ/NO

h) Integración en el entorno del usuario:

- ¿Permite la exportación de los resultados a otros formatos informáticos? **En caso negativo ponga NO**

ÓPTIMO/BUENO/REGULAR/MALO

- ¿Existen mensajes de ayuda sobre los resultados? SÍ/NO

- ¿Las direcciones web o de correo electrónico ¿se muestran como enlaces? SÍ/NO

2) Valoración sobre las secciones del periódico

a) Fidelidad expresiva:

- ¿Ha recuperado algún documento que no se corresponda con la categoría suscrita?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿Ha encontrado dos o más categorías o subcategorías que representen el mismo concepto?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿Podría incluir una nueva categoría o subcategoría en la lista de categorías?

ÓPTIMO/BUENO/REGULAR/MALO

b) Objetividad:

- ¿Conoce las variables que se utilizan para asignar las categorías?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿El sistema proporciona el listado de categorías y subcategorías? SÍ/NO

c) Pertinencia:

- ¿Cree que falta alguna categoría? ÓPTIMO/BUENO/REGULAR/MALO

- ¿Las categorías que el sistema le ofrece son las adecuadas?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿Cree que habrá documentos relevantes en otras categorías y que podrían aparecer en la suya? **En este caso ÓPTIMO indica que el sistema funciona bien, es decir, que no habrá noticias relevantes para usted en otras secciones y que podrían aparecer en la/s que ha seleccionado**

ÓPTIMO/BUENO/REGULAR/MALO

3) Valoración sobre las categorías (si las ha utilizado)

a) Fidelidad expresiva:

- ¿Ha recuperado algún documento que no se corresponda con la categoría suscrita?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿Ha encontrado dos o más categorías o subcategorías que representen el mismo concepto?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿Podría incluir una nueva categoría o subcategoría en la lista de categorías?

ÓPTIMO/BUENO/REGULAR/MALO

b) Objetividad:

- ¿Conoce las variables que se utilizan para asignar las categorías?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿El sistema proporciona el listado de categorías y subcategorías? SÍ/NO

c) Pertinencia:

- ¿Cree que falta alguna categoría? ÓPTIMO/BUENO/REGULAR/MALO

- ¿Las categorías que el sistema le ofrece son las adecuadas?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿Cree que habrá documentos relevantes en otras categorías y que podrían aparecer en la suya? **En este caso ÓPTIMO indica que el sistema funciona bien, es decir, que no habrá noticias relevantes para usted en otras categorías y que podrían aparecer en la/s que ha seleccionado**

ÓPTIMO/BUENO/REGULAR/MALO

4) Valoración sobre los resúmenes

a) Contenido del resumen:

- ¿Encuentra redundancias en el resumen? ÓPTIMO/BUENO/REGULAR/MALO

- ¿El resumen incluye valoraciones subjetivas**? En este caso, ÓPTIMO indica que no lo es en absoluto**. ÓPTIMO/BUENO/REGULAR/MALO

- ¿El resumen está recargado con detalles de interés secundario? ÓPTIMO/BUENO/REGULAR/MALO

- ¿El resumen se adapta al nivel de especialidad y al perfil del usuario? ÓPTIMO/BUENO/REGULAR/MALO

b) Estructuras del resumen:

- ¿Están representados los principales componentes de la noticia? ÓPTIMO/BUENO/REGULAR/MALO

- ¿Son correctas las frases del resumen? ÓPTIMO/BUENO/REGULAR/MALO

5) Medida de la relevancia de las noticias:

- Número de noticias solicitados diariamente:

- Número de noticias que NO ha solicitado y SÍ ha recibido (estimación aproximada diaria):

- Número de noticias (en los días de comprobación exhaustiva) que ha encontrado en el periódico y NO ha recibido:

(La relevancia de las noticias se mide tanto desde una perspectiva general como para permitir determinar si una noticia es la deseada o no)

- ¿Las noticias se corresponde con su perfil de interés?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿El contenido de la noticia tiene suficiente calidad en función de sus intereses?

ÓPTIMO/BUENO/REGULAR/MALO

B) PREGUNTAS ABIERTAS

1) ¿Cuáles son las características que consideras (5-10) más importantes del sistema?

2) ¿El sistema personaliza bien la información?

3) ¿Qué elementos echa en falta?

4) Desde su punto de vista, ¿el sistema es atractivo?

5) ¿Cree que permite y/o potencia la interactividad con el usuario?

6) ¿Cree que el sistema de categorías/secciones es bueno, porqué?

7) ¿Entiende cómo el sistema selecciona las noticias que usted quiere a partir de su perfil?

C) COMENTARIOS **(si usted lo desea, puede hacer cualquier tipo de comentario sobre el sistema, sobre todo en el modo de selección de las noticias)**

I.2. Cuestionario de evaluación para el sistema de personalización multilingüe

Este cuestionario contiene información sobre las opiniones de los usuarios. Es un documento Word con el siguiente aspecto:

Analista:

Navegador utilizado:

Fecha del análisis (días en los que ha usado el servicio):

Incidencias:

1) Evaluación general

a) Acceso:

- ¿El sistema es fácil de instalar, usar y mantener?

ÓPTIMO/BUENO/REGULAR/MALO

b) Interfaz:

- ¿Cuál es el grado de satisfacción con los componentes gráficos: iconos o barras de herramientas que permitan el acceso a todas las funciones?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿Existe alguna forma de interrumpir un proceso ya iniciado (búsqueda, alta o baja, etc.?

SÍ/NO

c) Adaptación al usuario:

- ¿Cómo es de amigable darse de alta y de baja en la suscripción? ÓPTIMO/BUENO/REGULAR/MALO

- ¿Existe un tutorial o una introducción al sistema? SÍ/NO

- Valore la ayuda al usuario que proporciona el sistema. ÓPTIMO/BUENO/REGULAR/MALO

d) Gestión de contenidos:

- ¿El sistema permite seleccionar varias categorías o secciones para confeccionar el perfil del usuario? SÍ/NO

- ¿Permite elegir los días de la semana en que se quiere recibir la información? SI/NO

- Valore el sistema de enlaces (colores, longitud, título, etc.)

ÓPTIMO/BUENO/REGULAR/MALO

- ¿Se resaltan los elementos consultados con anterioridad? SÍ/NO

- En qué medida cree que funciona el sistema de retroalimentación (sistema que permite mejorar los resultados conforme a la satisfacción del usuario con cada noticia recibida)

ÓPTIMO/BUENO/REGULAR/MALO

e) Esquemas de búsqueda:

- ¿Existe un sistema de búsqueda que le permita obtener noticias en los periódicos?

SI/NO

- ¿Permite el sistema que el usuario modifique sus criterios de búsqueda? SÍ/NO

- ¿El sistema ofrece el número de documentos incluidos en la recuperación? SÍ/NO

- ¿El sistema ofrece posibilidades de limitación en el número de documentos recuperados?

SÍ/NO

- ¿Aparece en las noticias recuperadas su grado de relevancia? SÍ/NO

f) Esquemas de recuperación y consulta:

- ¿Ordena los resultados por diferentes criterios (relevancia, por campos, etc.)? SÍ/NO

- ¿Permite acceder a los documentos desde cualquier sitio que se mencione? SÍ/NO

- Valore el esfuerzo y el tiempo requerido para el uso del sistema y para recibir las noticias

ÓPTIMO/BUENO/REGULAR/MALO

- ¿Permite el sistema el acceso al documento completo? SÍ/NO

- ¿En cada noticia se puede ver la sección, categoría o término al que está vinculada?

SÍ/NO

g) Integración en el entorno del usuario:

- ¿Las direcciones web o de correo electrónico ¿se muestran como enlaces? SÍ/NO

- ¿Es posible su envío por correo electrónico a otras personas? SÍ/NO

2) Valoración sobre las secciones y sobre las categorías del periódico

a) Fidelidad expresiva:

- Valore la calidad en la recuperación de documentos respecto a las categorías y secciones suscritas:

ÓPTIMO/BUENO/REGULAR/MALO

- ¿Ha encontrado dos o más categorías o secciones que representen el mismo concepto? SÍ/NO

- ¿En qué medida considera que la lista de categorías o secciones es completa? ÓPTIMO/BUENO/REGULAR/MALO

b) Objetividad:

- ¿El sistema proporciona directamente el listado de secciones? SÍ/NO

- ¿El sistema proporciona directamente el listado de categorías? SÍ/NO

- Valore sus conocimientos sobre las variables que se utilizan para asignar las noticias a una categoría o sección determinada

ÓPTIMO/BUENO/REGULAR/MALO

c) Pertinencia:

- ¿En qué medida cree que las categorías o secciones que el sistema ofrece son las adecuadas?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿Cree que habrá documentos relevantes en otras categorías o secciones y que podrían aparecer en las que usted ha elegido? SÍ/NO

d) Organización

- Señale el modo de perfil del usuario que le parezca más adecuado:

Secciones/Categorías/Términos

3) Valoración sobre los resúmenes

a) Contenido del resumen:

- ¿Encuentra redundancias en el resumen? SÍ/NO

- ¿En el resumen hay datos que no figuran en el documento original? SÍ/NO

- ¿El resumen está recargado con detalles de interés secundario? SÍ/NO

- ¿En qué medida el resumen se adapta al perfil del usuario? ÓPTIMO/BUENO/REGULAR/MALO

b) Estructuras del resumen:

- ¿En qué medida están representados los principales componentes de la noticia (Quién, Qué, Cómo, Cuándo, Dónde)?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿En qué medida el contenido del resumen está bien expresado (sintaxis, gramática, coherencia ...)?

ÓPTIMO/BUENO/REGULAR/MALO

4) Medida sobre la capacidad bilingüe del sistema

- ¿El sistema permite seleccionar el idioma? SÍ/NO

- Valore cómo ha traducido el sistema los términos que usted ha seleccionado en su perfil de usuario

ÓPTIMO/BUENO/REGULAR/MALO

- En qué medida cree que las noticias recibidas en uno y otro idioma responden a los términos seleccionados

ÓPTIMO/BUENO/REGULAR/MALO

5) Medida de la relevancia de las noticias:

(Realizar sobre cada uno de los documentos finales analizados)

- ¿El contenido de la noticia tiene suficiente calidad en función de sus intereses?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿El contenido del documento era lo que esperaba?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿En qué grado la noticia presenta la perspectiva, la profundidad y el planteamiento oportuno, desde su punto de vista?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿En qué medida le ofrece la noticia nuevo conocimiento frente a otras noticias relacionadas?

ÓPTIMO/BUENO/REGULAR/MALO

- ¿En qué grado el documento le resulta cercano en un sentido amplio, concuerda con su punto de vista, le afecta emocional o geográficamente?

ÓPTIMO/BUENO/REGULAR/MALO

6) Observaciones

I.3. Cuestionarios de evaluación para el sistema de personalización 2.0.

Existen dos cuestionarios, uno de evaluación inicial y otro de evaluación final. Estos cuestionarios contienen información sobre las opiniones de los usuarios.

I.3.1. Cuestionario de evaluación inicial

Contiene información sobre las opiniones de los usuarios antes de empezar a utilizar el sistema. Es un formulario HTML con el siguiente aspecto:

DATOS DEL EVALUADOR

Nombre completo:

Login de usuario en el sistema:

Tipo de evaluador: Profesor

Área del evaluador: Informática

Navegador utilizado: Internet Explorer

CUESTIONARIO

1) Evaluación de la interfaz

a) Interfaz General:

- ¿Cuál es el grado de satisfacción con los componentes gráficos?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO
☐ MUY BAJO

- ¿En qué grado el sistema es atractivo?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO
☐ MUY BAJO

b) Usabilidad:
- ¿En qué medida el sistema es fácil de usar?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO
☐ MUY BAJO

- ¿En qué grado el sistema es amigable para el usuario?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO
☐ MUY BAJO

c) Gestión de contenidos:
- ¿En qué grado el sistema de enlaces es bueno? (colores, longitud, título, etc.)

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

d) Sistemas de ayuda al usuario:
- Valore la ayuda que proporciona el sistema al usuario.

☐ MUY BUENA ☐ BUENA ☐ REGULAR
☐ MALA ☐ MUY MALA

2) Valoración sobre las secciones

- ¿En qué grado las secciones que el sistema le ofrece son adecuadas a la hora de reflejar sus necesidades de información?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO
☐ MUY BAJO

- ¿Introduciría nuevas secciones para reflejar sus necesidades de información? ¿Cuántas?

☐ MUCHAS ☐ ALGUNAS ☐ POCAS
☐ NINGUNA

3) Valoración sobre las categorías

- ¿En qué grado las categorías que el sistema le ofrece son adecuadas a la hora de reflejar sus necesidades de información?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO
☐ MUY BAJO

- ¿Introduciría nuevas categorías para reflejar sus necesidades de información? ¿Cuántas?

☐ MUCHAS ☐ ALGUNAS ☐ POCAS
☐ NINGUNA

4) Valoración sobre las palabras clave

- ¿En qué grado la posibilidad de introducir palabras clave que el sistema le ofrece es adecuada a la hora de reflejar sus necesidades de información?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO
☐ MUY BAJO

5) Medida de la relevancia de las noticias:

- Señale cuáles de estos criterios son los que va a aplicar a la hora de decidir si una noticia es relevante o no:

☐ La perspectiva y el planteamiento ☐ La profundidad y la cantidad de información

☐ El estilo ☐ La novedad ☐ La utilidad ☐ La relación con su perfil de usuario

☐ La relación con sus necesidades de información y con sus temas de interés

☐ La capacidad para añadir nuevo conocimiento frente a otros documentos relacionados

☐ La cercanía y grado de motivación desde un punto de vista emocional

☐ La cercanía desde un punto de vista geográfico

☐ La cercanía y familiaridad con el contenido expuesto

☐ La cercanía y familiaridad con el lenguaje empleado

- ¿Cómo describe su nivel del interés por la información que va a recibir?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

6) Estimación global del sistema:

- ¿Cuál es el nivel de confianza que tiene en el sistema antes de utilizarlo?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué nivel el sistema puede resolver sus necesidades de información iniciales?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿Qué sistema prefiere para poder definir sus intereses?

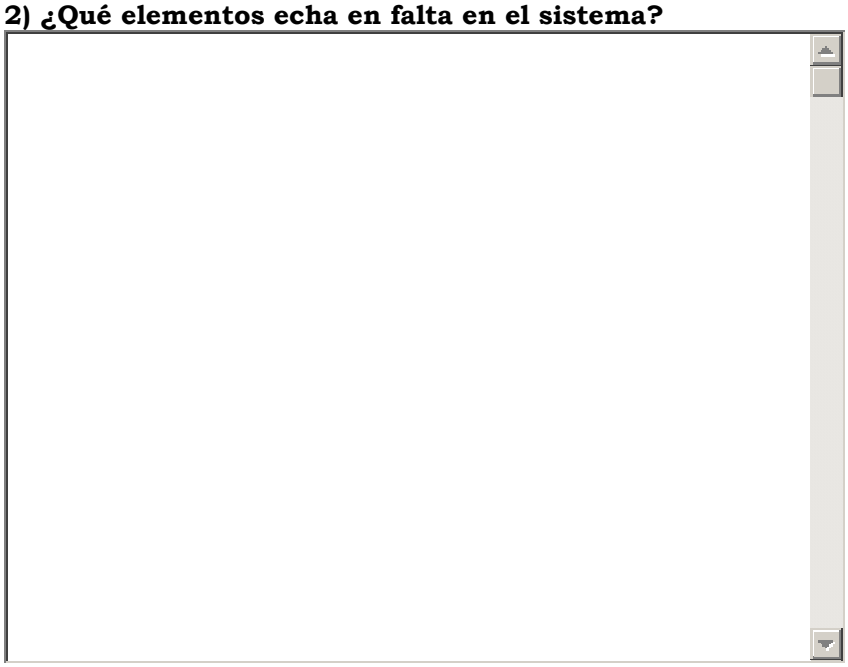
☐ SECCIONES ☐ CATEGORÍAS ☐ PALABRAS CLAVE ☐ OTROS

- En caso de otros, indique cuáles:

PREGUNTAS ABIERTAS

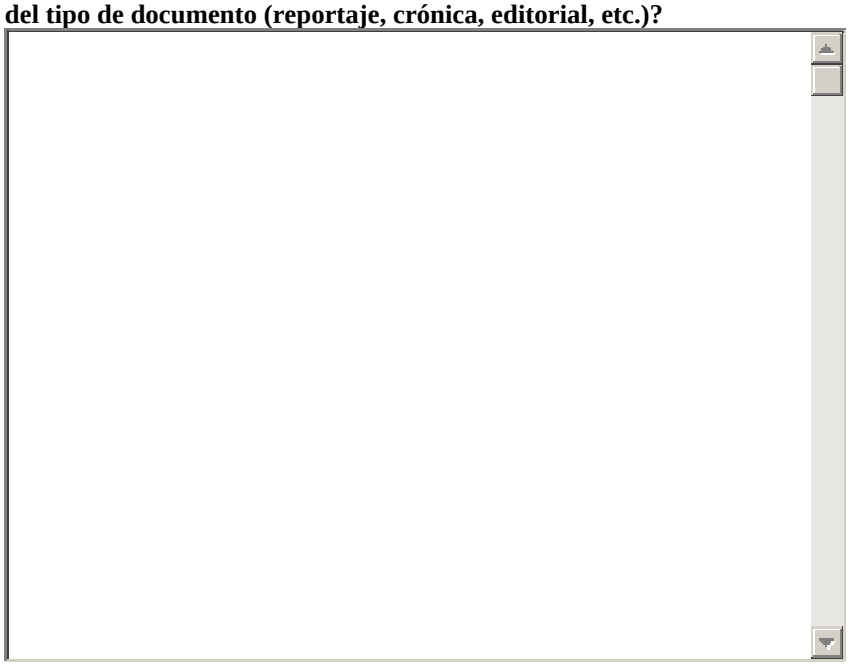
1) ¿Cuáles son las características que considera más importantes del sistema?

2) ¿Qué elementos echa en falta en el sistema?



3) Describa las necesidades de información que tiene antes de usar el sistema

4) ¿Qué tipo de información le interesa más, no sólo desde un punto de vista temático, sino del tipo de documento (reportaje, crónica, editorial, etc.)?



COMENTARIOS

Enviar

I.3.2. Cuestionario de evaluación final

Contiene información sobre las opiniones de los usuarios después de utilizar el sistema. Es un formulario HTML con el siguiente aspecto:

DATOS DEL EVALUADOR

Nombre completo:

Login de usuario en el sistema:

Profesión:

Edad:

Sexo: Masculino

Fechas de evaluación (días en los que ha usado el servicio):

Grado de experiencia en la utilización de sistemas de búsqueda similares:

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

CUESTIONARIO

1) Evaluación de la interfaz

a) Interfaz General:

- ¿Cuál es el grado de satisfacción con los componentes gráficos?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué grado el sistema es atractivo?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

b) Usabilidad:

- ¿En qué medida el sistema es fácil de usar?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué grado el sistema es amigable para el usuario?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

c) Gestión de contenidos:

- ¿En qué grado el sistema de enlaces es bueno? (colores, longitud, título, etc.)

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

d) Sistemas de ayuda al usuario:

- Valore la ayuda que proporciona el sistema al usuario.

☐ MUY BUENA ☐ BUENA ☐ REGULAR ☐ MALA ☐ MUY MALA

2) Valoración sobre las secciones

¿En qué grado las secciones que el sistema le ofrece han sido adecuadas a la hora de reflejar sus necesidades de información?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿Introduciría nuevas secciones para reflejar sus necesidades de información? ¿Cuántas?

☐ MUCHAS ☐ ALGUNAS ☐ POCAS ☐ NINGUNA

- Si añadiera otras secciones, ¿cuáles serían?:

- ¿En qué grado el sistema le muestra documentos correspondientes a las secciones escogidas antes que documentos que no pertenezcan a ellas?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿Ha cambiado durante el uso del sistema de secciones?

☐ SI ☐ NO

- En caso de cambios, indique los días en los que se realizaron

3) Valoración sobre las categorías

¿En qué grado las categorías que el sistema le ofrece han sido adecuadas a la hora de reflejar sus necesidades de información?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

¿Introduciría nuevas categorías para reflejar sus necesidades de información? ¿Cuántas?

☐ MUCHAS ☐ ALGUNAS ☐ POCAS ☐ NINGUNA

- Si añadiera otras categorías, ¿cuáles serían?:

- ¿En qué grado el sistema le muestra documentos correspondientes a las categorías escogidas antes que documentos que no pertenezcan a ellas?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿Ha cambiado durante el uso del sistema de categorías?

☐ SI ☐ NO

- En caso de cambios, indique los días en los que se realizaron

4) Valoración sobre las palabras clave

- ¿En qué grado la posibilidad de introducir palabras clave que el sistema le ofrece ha sido adecuada a la hora de reflejar sus necesidades de información?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué grado el sistema le muestra documentos correspondientes a las palabras clave escogidas antes que documentos que no las contengan?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué grado cree que los documentos recuperados según las palabras clave se corresponden con sus necesidades de información?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué grado quedan claras las razones por las que el sistema recupera los documentos a partir de las palabras propuestas?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué nivel los documentos recuperados se corresponden con el nivel de especificidad de las palabras propuestas?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué grado sería conveniente algún instrumento, como un diccionario, que le mostrara otras palabras relacionadas para mejorar la selección de las noticias, ?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿Ha cambiado durante el uso del sistema las palabras clave?

☐ SI ☐ NO

- En caso de cambios, indique los días en los que se realizaron

5) Valoración sobre los resúmenes:

- ¿En qué grado los resúmenes son de calidad?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué medida los resúmenes están bien construidos, son coherentes y claros?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué grado el sistema es capaz de evitar que se encuentren redundancias en el resumen?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué grado se adaptan los resúmenes a su perfil de usuario?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué grado se adaptan los resúmenes a sus necesidades de información?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué grado cree que los resúmenes reflejan el contenido de los documentos?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿Están representados en el resumen los principales componentes de la noticia?

☐ SI ☐ NO

- En caso de contestación negativa, indique qué componentes no están representados

6) Valoración sobre la selección y la adaptación:

- ¿En qué medida el sistema muestra las noticias correspondientes a sus necesidades de información antes que otras noticias?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué medida el sistema se adapta con el tiempo a sus necesidades de información?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué medida el sistema se adapta a los juicios que usted realiza sobre las noticias?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué medida han cambiado sus necesidades de información a lo largo de la utilización del sistema?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

7) Medida de la relevancia de las noticias:

- Señale cuáles de estos criterios son los que ha aplicado a la hora de decidir si una noticia era relevante o no:

☐ La perspectiva y el planteamiento ☐ La profundidad y la cantidad de información

☐ El estilo ☐ La novedad ☐ La utilidad ☐ La relación con su perfil de usuario

☐ La relación con sus necesidades de información y con sus temas de interés

☐ La capacidad para añadir nuevo conocimiento frente a otros documentos relacionados

☐ La cercanía y grado de motivación desde un punto de vista emocional

☐ La cercanía desde un punto de vista geográfico

☐ La cercanía y familiaridad con el contenido expuesto

☐ La cercanía y familiaridad con el lenguaje empleado

- ¿Cómo describe su nivel de interés por la información que ha recibido, después de usar el sistema?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿Señale cuántas veces ha usado cada una de estas informaciones relacionadas con cada noticia para decidir si una noticia era relevante o no?

- Titular:

☐ MUCHAS ☐ ALGUNAS ☐ POCAS
☐ NINGUNA

- Sección:

☐ MUCHAS ☐ ALGUNAS ☐ POCAS
☐ NINGUNA

- Relevancia:

☐ MUCHAS ☐ ALGUNAS ☐ POCAS
☐ NINGUNA

- Resumen:

☐ MUCHAS ☐ ALGUNAS ☐ POCAS
☐ NINGUNA

- Noticia completa:

☐ MUCHAS ☐ ALGUNAS ☐ POCAS
☐ NINGUNA

6) Estimación global del sistema:

¿Cuál es el nivel general de satisfacción y confianza que tiene en el sistema, después de utilizarlo?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

- ¿En qué grado el sistema ha resuelto sus necesidades de información?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

¿Qué sistema prefiere para poder definir sus intereses?

☐ SECCIONES ☐ CATEGORÍAS ☐ PALABRAS CLAVE ☐ OTROS

- En caso de otros, indique cuáles:

- ¿En qué medida el sistema personaliza bien la información?

☐ MUY ALTO ☐ ALTO ☐ REGULAR ☐ BAJO ☐ MUY BAJO

PREGUNTAS ABIERTAS

1) ¿Cuáles son las características que considera más importantes del sistema?

2) ¿Qué elementos echa en falta en el sistema?

3) Describa las necesidades de información que tiene después de usar el sistema

4) ¿Cree que permite la interactividad con el usuario?

5) ¿Qué tipo de información le ha interesado más, no sólo desde un punto de vista temático, sino del tipo de documento (reportaje, crónica, editorial, etc.), después de usar el sistema?

6) Si ha cambiado su perfil de usuario, ¿cuáles han sido las razones?

COMENTARIOS

Enviar

Apéndice II

EJEMPLOS DE NOTICIA Y MODELO DE USUARIO

En este apéndice se va a mostrar un ejemplo de una noticia y de un modelo de usuario utilizado en el sistema de personalización 1.0. Las noticias utilizadas en los otros experimentos tienen un formato y una longitud similar. En cuanto a los modelos de usuario, en el segundo sistema también aparecen las categorías como sistema de referencia para el modelo a largo plazo. Por otro lado, en el sistema multilingüe hay noticias en español y en inglés y el modelo de usuario contiene información sobre secciones, categorías y palabras clave, en ambos idiomas.

La noticia corresponde a la sección de nacional y al día 8 de Febrero de 2002.

Ejemplo de noticia.

Detenidos cuatro etarras que ayudaban a pasar la frontera a los «liberados»

D. M. / J. P.

El ministro del Interior, Mariano Rajoy, destacó ayer que la operación policial es fruto de varios meses de investigación de la Guardia Civil, cuyo punto de partida fue una pista obtenida en Francia. Concretamente, algunos de los ahora detenidos fueron identificados en el país vecino cuando ayudaban a miembros «liberados» (fichados por las Fuerzas de Seguridad) de la banda a cruzar la frontera con España. Esta era el principal cometido de los cuatro arrestados: Miren Agurtzane Uriarte Bustinza, de 34 años; Patxi Xabier Ascaso Errasti, de 36; José Ramón Revilla Arbaiza y Aritz Zabaleta Jáuregui, de 26 años. Los tres primeros fueron detenidos en la localidad vizcaína de Lequeitio y el último en Bilbao.

Además, la operación, dirigida por el juez de la Audiencia Nacional Luis del Olmo, se saldó con el registro de tres viviendas, situadas en las calles Letraukua y Aguirre Solarte en Lequeitio y en la de Iturribide de Bilbao, y con la incautación de diversa documentación que se está analizando. A ninguno de los detenidos se les ocupó armas.

«Lanzadera» y enlace

El ministro del Interior afirmó que los cuatro etarras servían presuntamente de enlace en Francia entre los miembros de la banda. Concretamente, precisó que una de las funciones de los arrestados era hacer de «lanzadera» para la entrada en España de miembros «liberados» de ETA. Esta función consiste en ir por delante de los vehículos en los que se trasladan los «comandos» o los explosivos para avisar si se encuentran en el trayecto con un control policial, por lo que los terroristas suelen ir conectados con teléfonos móviles.

Mariano Rajoy también señaló que los arrestados habían mantenido «entrevistas» en territorio francés con miembros de ETA, aunque no quiso revelar ni cuando ni con quienes.

Aunque las investigaciones siguen abiertas, de momento no se cree que los detenidos hayan participado en los últimos atentados cometidos en la provincia de Vizcaya.

Según el ministro del Interior, todos ellos han estado vinculados con los grupos de apoyo a la banda. La de ayer no era la primera detención de Miren Agurtzane Uriarte. El 27 de abril de 1988 fue arrestada también por la Guardia Civil durante una operación contra la infraestructura de ETA, pero al día siguiente fue puesta en libertad.

El modelo de usuario también corresponde al día 8 de Febrero de 2002, por lo tanto, tiene los términos de realimentación procedentes de la realimentación producida los días 6 y 7 de Febrero. Entre paréntesis aparecen los pesos asociados a cada uno de los elementos del modelo.

Ejemplo de modelo de usuario.

Secciones: nacional (1) y deportes (1).

Palabras clave: Gil (1), atlético (1), ETA (1).

Términos de realimentación: Batasuna (0.918), Madrid (0.693), Abaúnza (0.586), Francia (0.537), fiscalía (0.533), extradición (0.533), red (0.533), Martínez (0.514), proceso (0.470), comisiones (0.427), atlético (0.567).

Apéndice III
EJEMPLOS DE RESÚMENES GENERADOS

En este apéndice se va a mostrar un ejemplo de distintos resúmenes generados para una misma noticia utilizando un modelo de usuario concreto. En particular, se va a utilizar la noticia y el modelo de usuario presentados en el apéndice anterior. Los resúmenes para otras noticias tendrán un aspecto similar, aunque cuanto más larga sea la noticia más largo será el resumen. En el modelo multilingüe habrá resúmenes en inglés y en español correspondientes a noticias en inglés y en español, respectivamente.

Resumen base.

Detenidos cuatro etarras que ayudaban a pasar la frontera a los «liberados»

D. M. / J. P.

El ministro del Interior, Mariano Rajoy, destacó ayer que la operación policial es fruto de varios meses de investigación de la Guardia Civil, cuyo punto de partida fue una pista obtenida en Francia. Concretamente, algunos de los ahora detenidos fueron identificados en el país vecino cuando ayudaban a miembros «liberados» (fichados por las Fuerzas de Seguridad) de la banda a cruzar la frontera con España. Esta era el principal cometido de los cuatro arrestados: Miren Agurtzane Uriarte Bustinza, de 34 años; Patxi Xabier Ascaso Errasti, de 36; José Ramón Revilla Arbaiza y Aritz Zabaleta Jáuregui, de 26 años.

Resumen genérico.

Detenidos cuatro etarras que ayudaban a pasar la frontera a los «liberados»

D. M. / J. P.

Concretamente, algunos de los ahora detenidos fueron identificados en el país vecino cuando ayudaban a miembros «liberados» (fichados por las Fuerzas de Seguridad) de la banda a cruzar la frontera con España. Esta era el principal cometido de los cuatro arrestados: Miren Agurtzane Uriarte Bustinza, de 34 años; Patxi Xabier Ascaso Errasti, de 36; José Ramón Revilla Arbaiza y Aritz Zabaleta Jáuregui, de 26 años. Los tres primeros fueron detenidos en la localidad vizcaína de Lequeitio y el último en Bilbao.

Resumen personalizado.

Detenidos cuatro etarras que ayudaban a pasar la frontera a los «liberados»

D. M. / J. P.

El ministro del Interior afirmó que los cuatro etarras servían presuntamente de enlace en Francia entre los miembros de la banda. Concretamente, precisó que una de las funciones de los arrestados era hacer de «lanzadera» para la entrada en España de miembros «liberados» de ETA.

Mariano Rajoy también señaló que los arrestados habían mantenido «entrevistas» en territorio francés con miembros de ETA, aunque no quiso revelar ni cuando ni con quienes.

Resumen genérico-personalizado.

Detenidos cuatro etarras que ayudaban a pasar la frontera a los «liberados»

D. M. / J. P.

Concretamente, algunos de los ahora detenidos fueron identificados en el país vecino cuando ayudaban a miembros «liberados» (fichados por las Fuerzas de Seguridad) de la banda a cruzar la frontera con España.

El ministro del Interior afirmó que los cuatro etarras servían presuntamente de enlace en Francia entre los miembros de la banda.

Mariano Rajoy también señaló que los arrestados habían mantenido «entrevistas» en territorio francés con miembros de ETA, aunque no quiso revelar ni cuando ni con quienes.

Apéndice IV
EJEMPLO DE PÁGINA DE YAHOO!

En este apéndice se va a mostrar un ejemplo de las páginas asociadas a las categorías de Yahoo!. Hay que tener en cuenta que las páginas utilizadas fueron las que existían en Yahoo! cuando se desarrolló el primer experimento preliminar, esto es, principios del 2000. En el sistema multilingüe se utilizan además las páginas de Yahoo! Estados Unidos.

El ejemplo mostrado es la página de primer nivel de la categoría Education de Yahoo! Estados Unidos. Sólo se muestra la parte que se utiliza para construir la representación de las categorías. Todas las páginas de Yahoo!, estén en el idioma que estén, tienen el mismo formato.

- **Academic Competitions** *(71)*
- **Adult and Continuing Education** *(290)* NEW!
- **Bibliographies** *(5)*
- **Bilingual** *(17)*
- **Career and Vocational** *(223)* NEW!
- **Chats and Forums** *(46)*
- **Companies@**
- **Conferences** *(39)* NEW!
- **Correctional@**
- **Disabilities@**
- **Distance Learning** *(422)* NEW!
- **Early Childhood Education** *(72)* NEW!
- **Employment** *(122)* NEW!
- **Equity** *(26)*
- **Financial Aid** *(367)* NEW!
- **Government Agencies** *(74)*
- **Graduation** *(57)*
- **Higher Education** *(12987)* NEW!
- **Instructional Technology** *(321)* NEW!
- **Journals** *(27)*
- **K-12** *(44962)* NEW!
- **Literacy** *(9)*
- **News and Media** *(79)*
- **Organizations** *(2850)* NEW!
- **Policy** *(44)*
- **Programs** *(270)* NEW!
- **Reform** *(45)*
- **Special Education** *(153)*
- **Standards and Testing** *(61)*
- **Statistics** *(7)*
- **Teaching** *(89)*
- **Theory and Methods** *(560)* NEW!
- **Web Directories** *(41)*

- **ERIC@**
- Eurydice - information network on education in Europe.
- State of American Education - speech by U.S. Secretary of Education Richard W. Riley. Feb. 22, 2000.
- Study Guides and Strategies

Figura IV.1. Ejemplo de página de Yahoo!

Apéndice V

ESQUEMA DE LA BASE DE DATOS DE LOS SISTEMAS DE PERSONALIZACIÓN

En este apéndice se va a mostrar el esquema de la base de datos utilizado en la última versión de sistema de personalización (Figura V.I). Este esquema ha dado lugar a las siguientes 13 tablas: usuario, sección, categoría, catpersonal, usuariosección, usuariocategoría, usuariocat-personal, término, trealiment, noticia, usuarionoticia, catgeneral y noticiacatgeneral.

La base de datos utilizada en las otras versiones de sistemas de personalización varía en la ausencia de categorías, resúmenes o términos de realimentación. En el sistema multilingüe el esquema está duplicado, de tal forma que hay una versión de tabla por idioma, excepto de la tabla usuario, que además almacena el idioma seleccionado por el propio usuario.

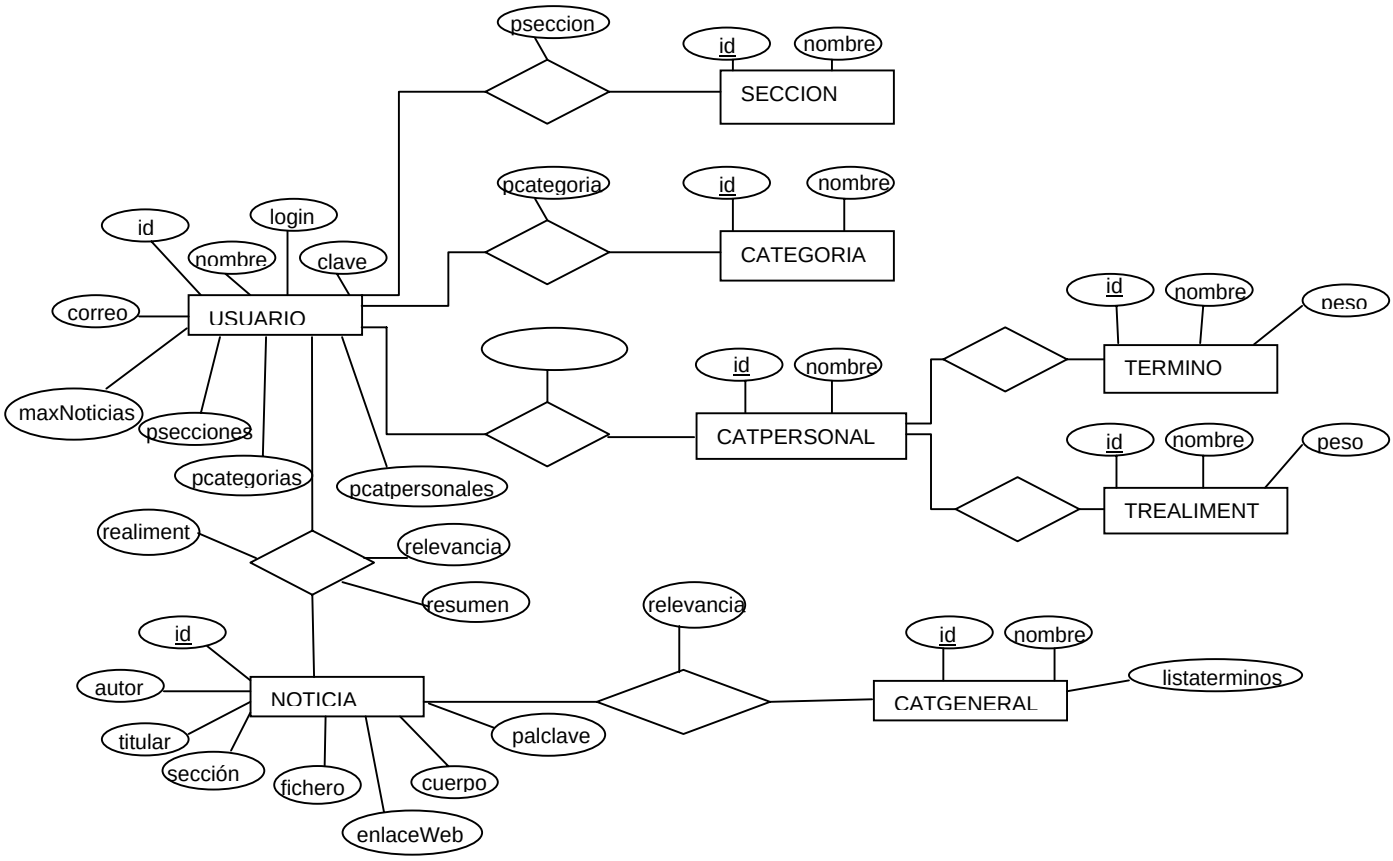


Figura V.1. Esquema de la base de datos

Apéndice VI

MANUAL DE USUARIO DE LOS SISTEMAS DE PERSONALIZACIÓN

A continuación se va a presentar el manual de usuario de los diferentes sistemas de personalización presentados en esta tesis. En realidad, la forma de manejar los sistemas es muy similar, puesto que todos están basados en la misma propuesta de personalización. La principal diferencia estriba en que en los experimentos preliminares no había realimentación y que en el sistema multilingüe se manejaba información en dos idiomas, aunque el usuario sólo tenía que rellenar el modelo de usuario en el idioma que él seleccionaba.

Lo que se va a presentar a continuación es un manual de usuario para el último sistema de personalización de noticias.

Cuando un usuario se conecta al sistema se le presenta la página de inicio del sistema (Figura VI.1). En esta página se presenta una pequeña introducción del sistema y se permite el acceso a las siguientes funcionalidades: Inicio de sesión, Alta de usuario, Modificar Perfil, Baja Usuario. Además el usuario puede acceder al cuestionario de evaluación final directamente desde esta página.

La primera vez que un usuario se conecta al sistema ha de darse de alta (Figura VI.2). Si, por el contrario, ya está conectado, bastará con que introduzca su login y clave y pulse el botón Iniciar. En todo caso, tras darse de alta o iniciar sesión, se mostrará la página de edición del modelo de usuario (Figuras VI.3, VI.4 y VI.5), donde el usuario podrá configurar sus intereses de acuerdo a los distintos sistemas de referencia: secciones, categorías y palabras clave. Cuando el usuario decida darse de baja lo realizará a través de la página correspondiente (Figura VI.6).

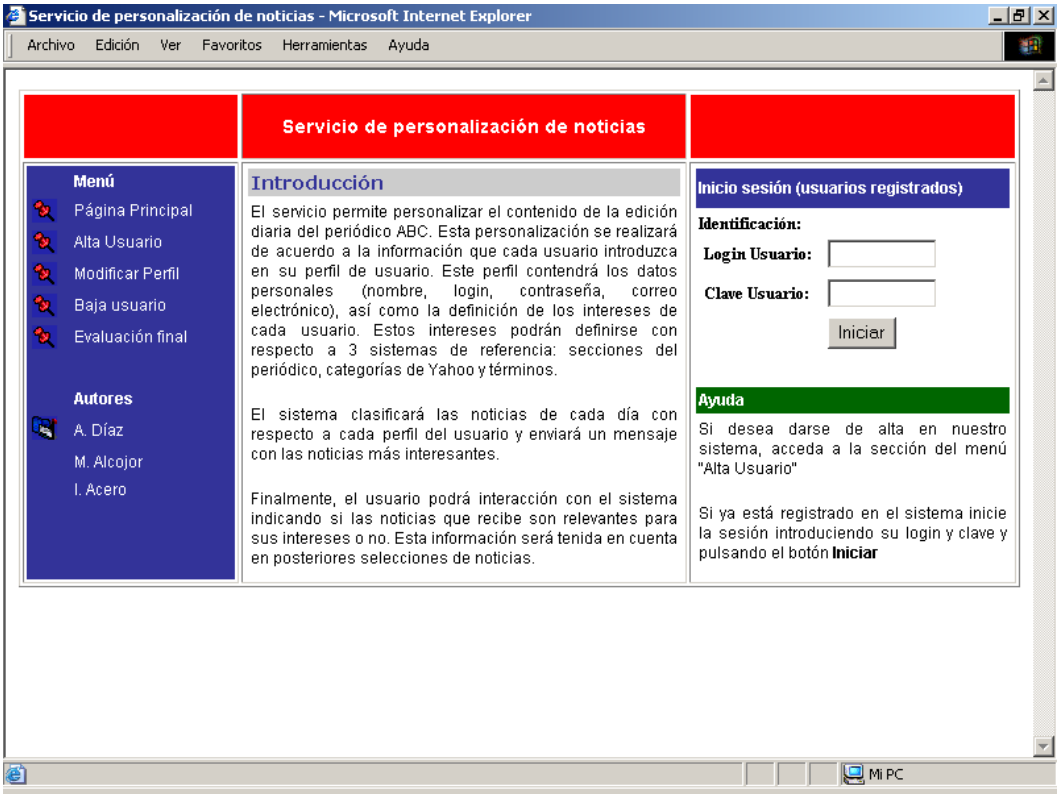


Figura VI.1. Página de inicio del sistema de personalización de noticias.

El proceso de alta (Figura VI.2) consiste en la introducción de una serie de datos personales asociados a los usuarios: nombre completo, correo electrónico, login y clave de usuario. Desde esta página existe un enlace al cuestionario de evaluación inicial que han de rellenar los usuarios justo después de registrarse. En este sistema no se introducía número máximo de noticias que el usuario quería recibir porque se le mandaban todas para poder construir la colección de evaluación.

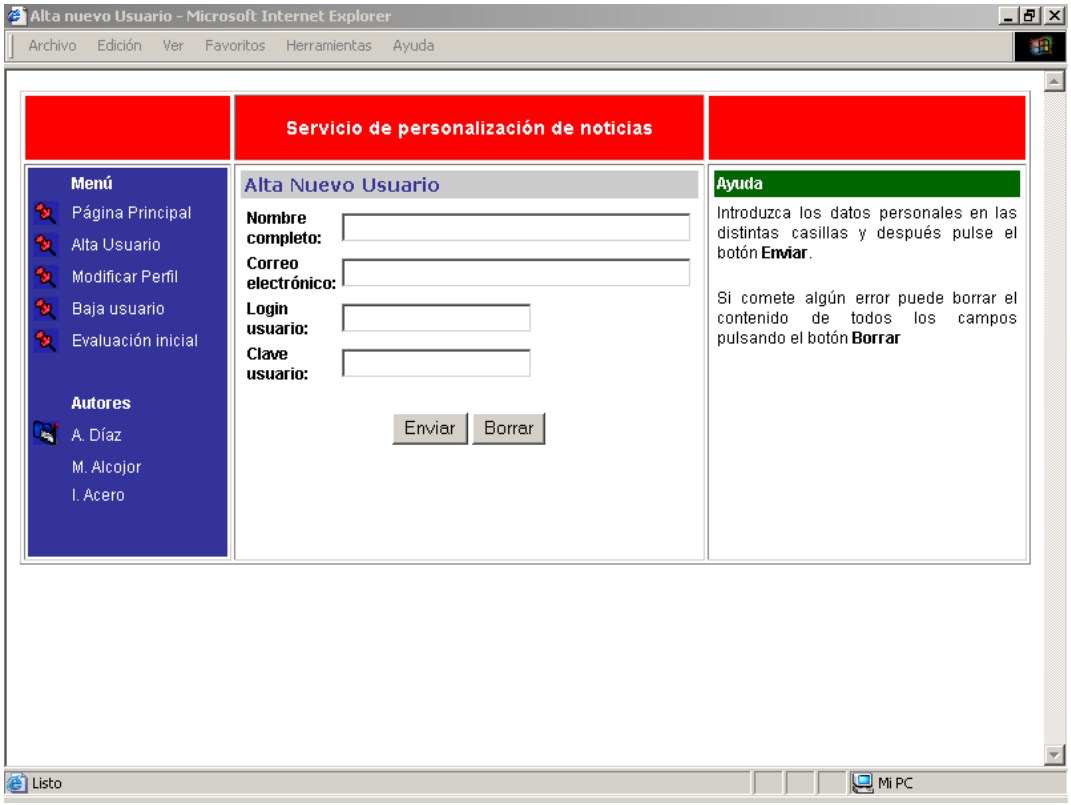


Figura VI.2. Página de alta.

En la edición del perfil hay que indicar inicialmente el interés por las secciones en general y posteriormente se indica el grado de interés en cada una de las secciones en particular. Las posibilidades son: Nada, Poco, Normal, Mucho (Figura VI.3).

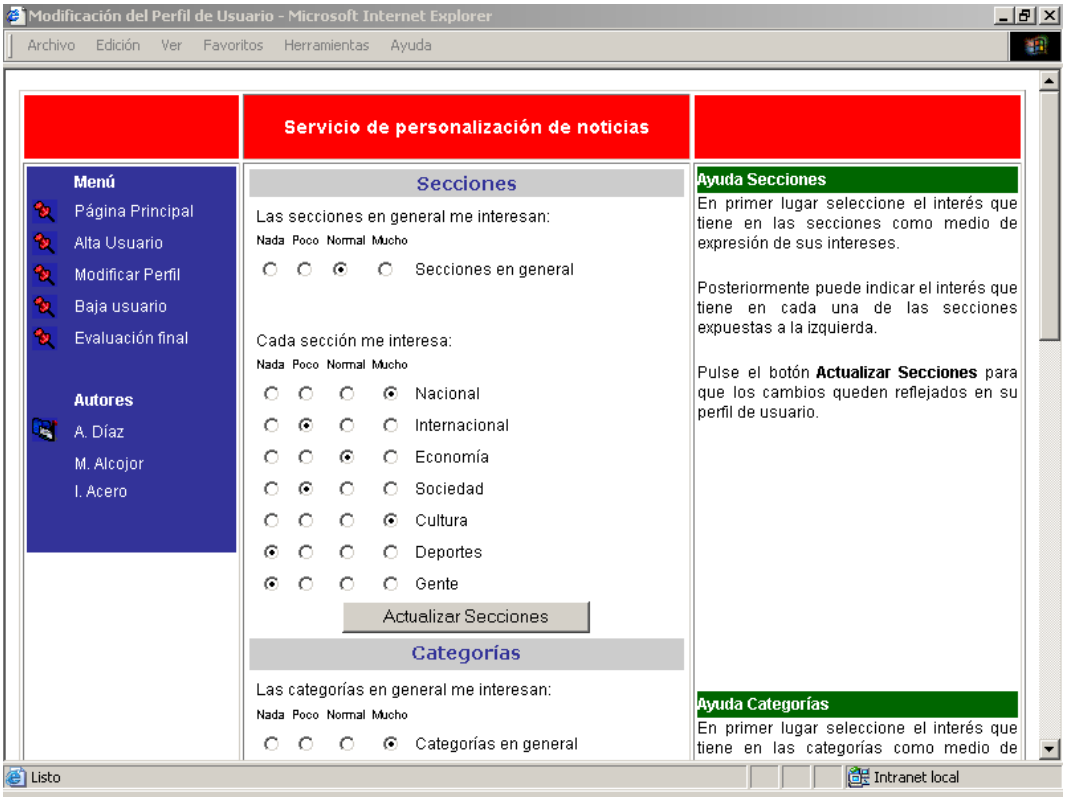


Figura VI.3. Edición del modelo de usuario (Secciones)

Para las categorías la estructura es básicamente igual que la usada para las secciones. Se debe elegir el interés general para las categorías y seguidamente seleccionar el interés para cada una de ellas (Figura VI.4).

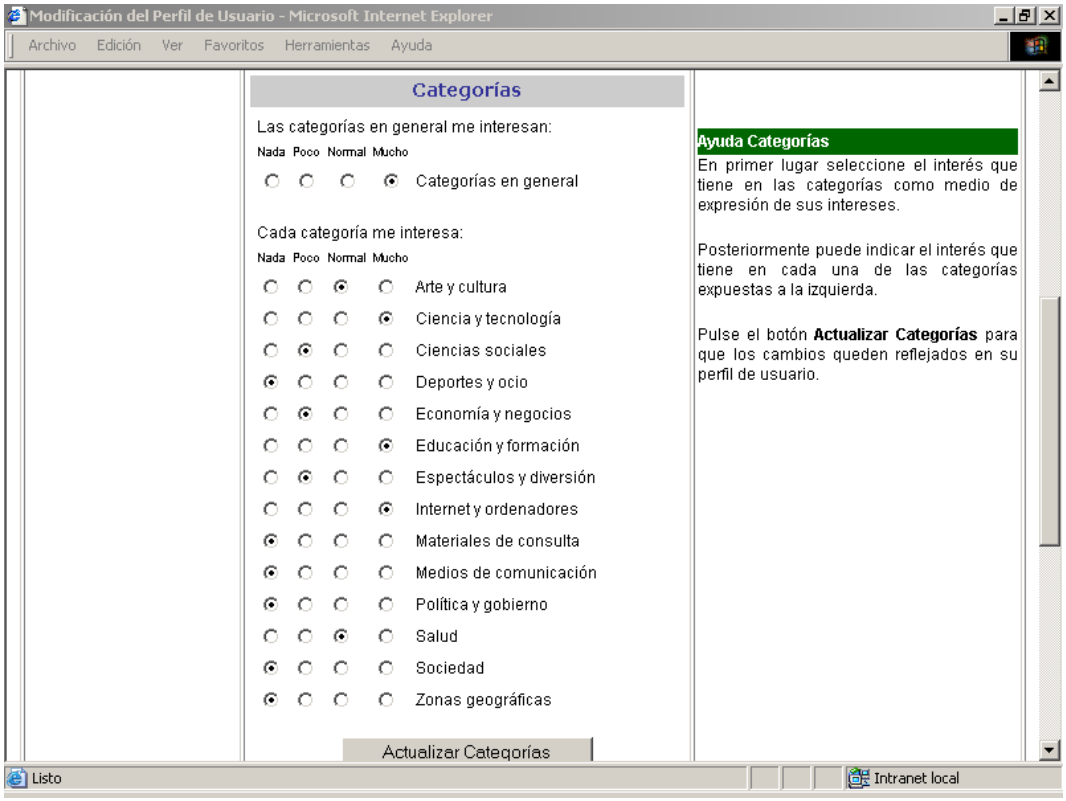


Figura VI.4. Edición del modelo de usuario (Categorías)

Por último, para las palabras clave o términos personales, también se dispone de un indicador de interés global para estos términos. El paso siguiente es escribir un término y añadirlo a la lista de términos definidos (si es que se ha definido algún otro). Se les puede asignar los grados de importancia: Alta, Media y Baja. Además tenemos 2 botones para cada término, con los que podemos borrar el término o modificarlo (Figura VI.5).

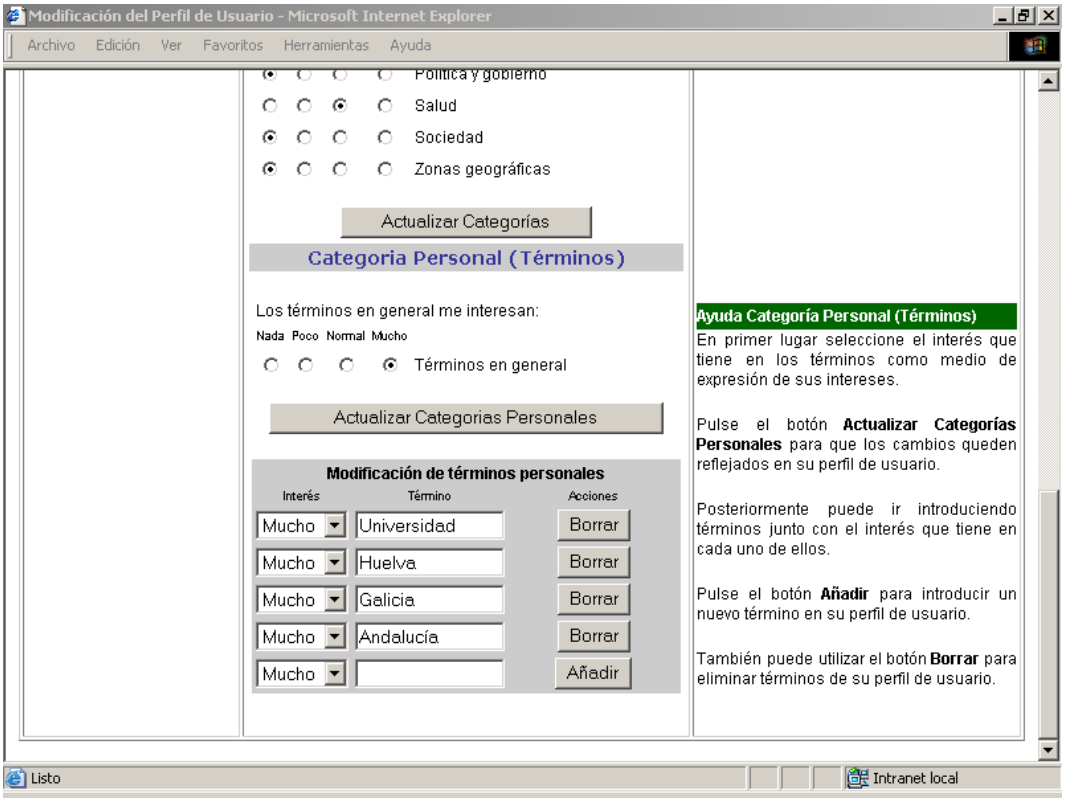


Figura VI.5. Edición del modelo de usuario (Palabras clave)

Para darse de baja (Figura VI.5) hay que indicar el login y la clave de usuario. Una vez hecho esto, toda la información que el sistema guardaba se borrará automáticamente.

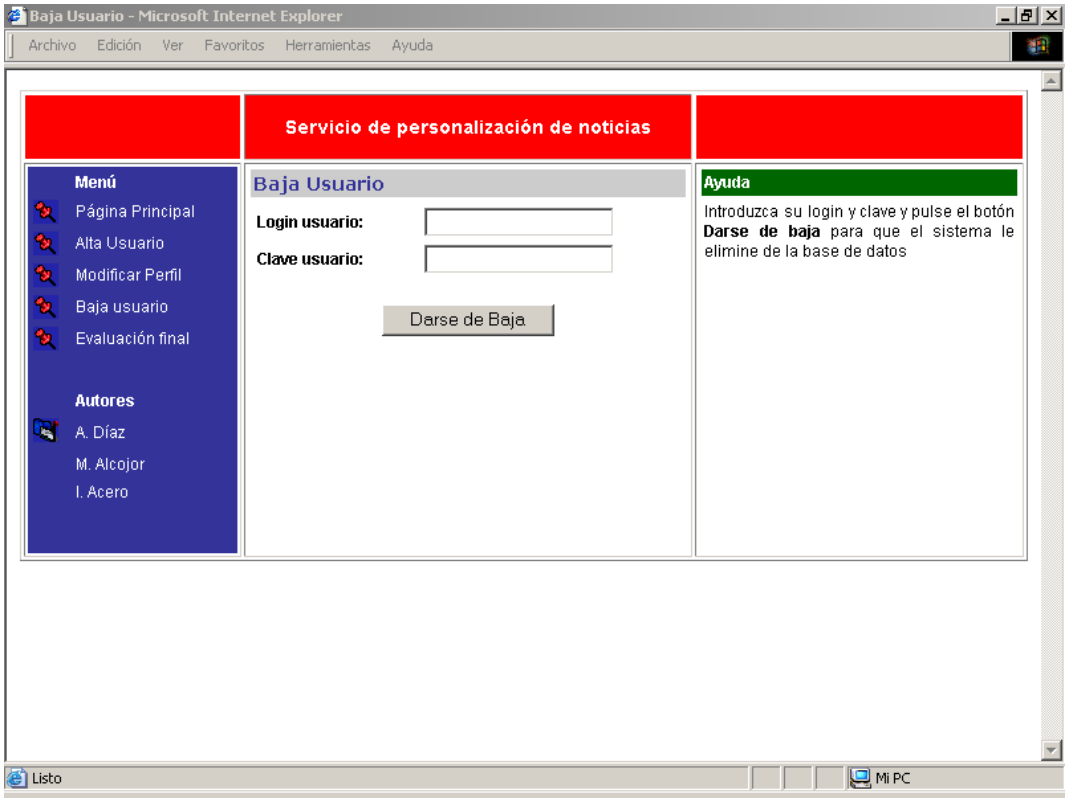


Figura VI.6. Baja de usuario

Con la intención de mejorar la interacción con el usuario, se ha incluido en la parte derecha de cada página, un apartado destinado a la ayuda sobre las diferentes acciones que puede realizar un usuario en cada página en particular. Se puede decir, pues, que la ayuda es sensitiva, respecto a la página en la que nos encontremos. De esta manera se consigue que el usuario se encuentre orientado en todo momento.

Por otro lado, el usuario recibe un mensaje como el que se presenta de ejemplo en la Figura VI.7:

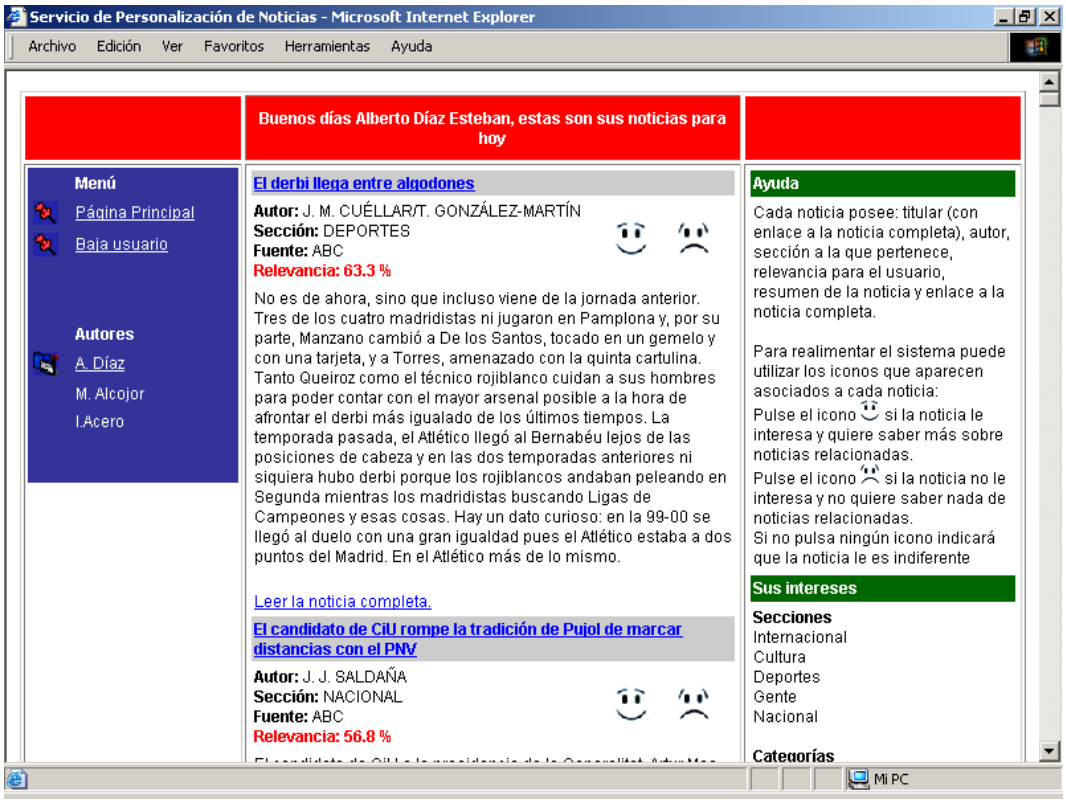


Figura VI.7. Ejemplo de mensaje.

El mensaje está personalizado con el nombre del usuario y contiene enlaces que permiten acceder a las distintas funcionalidades del sistema. Para cada noticia se incluye la siguiente información: título (con hiperenlace que permite acceder directamente a la noticia original), nombre de la sección del periódico a la que pertenece la noticia, fuente de la que procede, valor de relevancia calculado para la noticia en función de la información del perfil del usuario, resumen personalizado del contenido de la noticia, enlace adicional en que se menciona explícitamente que desde allí es posible acceder a la noticia completa e iconos de realimentación.

Las noticias se presentan ordenadas en función del valor de relevancia calculado. En la parte derecha del mensaje, debajo de la ayuda, se incluye un breve resumen de la caracterización de los intereses del usuario reflejados en su perfil.