

# PERSONALIZACIÓN EN SERVICIOS DE NOTICIAS

**Alberto Díaz Esteban**  
**Departamento de Ingeniería del Software e Inteligencia Artificial**  
**Universidad Complutense de Madrid**

## Resumen

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. La personalización de contenidos es una técnica que trata de eliminar la sobrecarga de información mediante la adaptación de los contenidos Web a cada tipo de usuario. En este trabajo 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 están reflejadas en un perfil o modelo de usuario que almacena intereses a largo y corto plazo. Para evaluar los distintos procesos de personalización se ha generado una colección de evaluación donde se almacenan los juicios de relevancia de varios usuarios durante varios días de utilización del sistema. 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 en un sistema de personalización de contenidos Web aplicado a servicios de noticias.

**Palabras clave:** personalización de contenidos, modelo de usuario, evaluación

## 1. Introducción

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. La personalización de contenidos es una técnica que trata de eliminar la sobrecarga de información mediante la adaptación de los contenidos Web a cada tipo de usuario.

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 (Díaz, 2005). Para que estas funcionalidades se realicen de manera personalizada deben ser realizadas en base a la información sobre los intereses del usuario, que están reflejados en su perfil o modelo de usuario. 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. 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 (Billsus&Pazzani, 2000). La presentación de resultados consiste en, una vez seleccionados los elementos de información que le interesan a un usuario, generar una presentación que contenga esos elementos de manera personalizada.

En este trabajo se presenta un modelo de personalización de contenidos Web aplicado al dominio de los servicios de noticias. El modelo de usuario propuesto integra varios sistemas de referencia que permiten representar los intereses de los usuarios desde diferentes puntos de vista: intereses a largo plazo e intereses a corto plazo. Los primeros representan intereses del usuario que permanecen constantes a lo largo del tiempo, los segundos representan los intereses que se van modificando a lo largo del tiempo.

Los resultados se muestran a los usuarios en forma de resúmenes personalizados. Esto 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.

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.

Por otro lado, en este trabajo se propone un marco de evaluación de sistemas de personalización de contenidos que permite establecer la efectividad del sistema (evaluación cuantitativa) y la satisfacción de los usuarios (evaluación cualitativa) cuando se utilizan distintas propuestas de personalización. Además se ha construido una colección de evaluación respecto a la cual se evalúan las técnicas de personalización presentadas.

El objetivo final del trabajo es mostrar 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 en un sistema de personalización de contenidos Web aplicado a servicios de noticias.

En el punto 2 se describe el estado del arte en personalización. La propuesta de personalización de contenidos se presenta en el apartado 3. La metodología de evaluación se describe en el punto 4 y los resultados obtenidos en el punto 5. Finalmente, se muestran las conclusiones obtenidas.

## **2. Estado del arte**

Existen en la bibliografía distintas técnicas para definir los intereses de los usuarios: términos, estereotipos, redes semánticas, redes neuronales, etc. En particular hay un conjunto de propuestas (Billsus&Pazzani, 2000; Widyantoro et al., 2001) que modelan a los usuarios combinando intereses a largo y corto plazo: el modelo a corto plazo representa las preferencias del usuario más recientes y el largo plazo las que se han producido en un período de tiempo más largo.

Para obtener una representación del contenido textual de los documentos se suelen utilizar técnicas basadas en vectores de pesos de términos. Los vectores asociados a los documentos se obtienen eliminando las palabras vacías almacenadas en una lista de parada y obteniendo las raíces de las palabras. Para calcular los pesos se suele utilizar fórmulas basadas en las frecuencias de aparición de los términos en los documentos (Salton&McGill, 1983).

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, reglas asociadas a estereotipos, redes neuronales, vecino más cercano, clasificador bayesiano ingenuo, etc.

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: actualización de pesos de términos, actualización de redes semánticas, actualización de reglas asociadas a estereotipos, etc.

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.

### **3. Personalización de contenidos Web**

En los siguientes puntos se describe el modelo de usuario, el proceso de selección utilizando el modelo de usuario completo y el proceso de presentación de resultados utilizando generación de resúmenes personalizados.

#### **3.1. Modelo de usuario**

El modelo de usuario propuesto (Díaz&Gervás, 2005) integra varios 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 y son definidos por el usuario en el momento de registrarse en el sistema. Los intereses a corto plazo representan los intereses que se van modificando a lo largo del tiempo y son obtenidos a partir del proceso de adaptación del modelo de usuario. Este proceso parte de la realimentación que el usuario efectúa sobre las noticias que recibe (me interesa / indiferente / no me interesa).

Para modelar los intereses a largo plazo se utilizan tres marcos de referencia: uno basado en un sistema de clasificación dependiente del dominio, otro basado en un algoritmo de categorización automática y otro basado en un conjunto de palabras claves.

Un sistema básico de referencia es el sistema de clasificación dependiente del dominio (p.ej. en un periódico electrónico, este sistema estaría basado en las secciones del periódico). Cada documento Web pertenece a una categoría de dicho sistema de clasificación. Los usuarios pueden asignar un peso a cada categoría para indicar su interés en cada una de ellas. El segundo sistema de referencia consiste en un conjunto de palabras clave que son introducidas por el usuario, con un peso asociado. El tercer sistema de referencia es un conjunto de categorías generales independientes del dominio utilizado. 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. La representación de las categorías generales se obtiene, en forma de vectores de pesos de términos, a partir de las páginas indexadas en el primer y segundo nivel de Yahoo! España.

Para representar los intereses a corto plazo se utilizan los términos de realimentación. Estos términos se obtienen de la realimentación del usuario sobre los documentos que recibe. Esto es, el usuario realimenta positiva o negativamente sobre los documentos recibidos y de ellos se extraen una serie de términos representativos. Esta información es manejada por el proceso de adaptación del modelo de usuario el cual devuelve un vector de pesos de términos para cada usuario, que representa este interés del usuario (Díaz&Gervás, 2005).

#### **3.2. Selección de contenidos**

Los documentos se bajan de la Web en forma de documentos HTML. Se extraen el título, categoría, URL y texto de cada documento y se almacenan para ser procesados posteriormente. Para obtener la representación de los documentos se utilizan vectores de pesos de términos aplicando lista de parada y extractor de raíces, y utilizando  $tf \cdot idf$  (Salton&McGill, 1983) para el cálculo de los pesos.

Como cada documento Web tiene una sección pre-asignada, la selección respecto este marco de referencia es inmediata. Se obtiene para cada documento el valor asignado a la sección a la que pertenece, dentro del modelo de usuario. La relevancia entre un documento y las palabras clave de un modelo de usuario se calcula mediante la fórmula de la similitud del coseno del modelo del espacio vectorial (Salton&McGill, 1983). De manera similar se calcula la relevancia entre un documento y los términos de realimentación del modelo a corto plazo. La similitud entre un documento y las categorías generales es calculada mediante la combinación lineal ponderada de las similitudes entre el documento y cada una de las categorías generales.

Cuando todos los documentos han sido ordenados con respecto a las diferentes fuentes de relevancia, los resultados son integrados utilizando una combinación lineal ponderada de las similitudes obtenidas. Los resultados de esta combinación establecen la similitud entre un documento y un modelo de usuario. El resultado final del proceso de selección es un ranking de documentos ordenado con respecto a este valor de similitud.

### **3.3. Presentación de resultados**

La presentación de resultados consiste en, una vez seleccionados los documentos que le interesan a un usuario, generar una presentación que contenga estos documentos como elementos de información.

El formato del documento Web generado es el siguiente: un título con la fecha y el nombre del usuario, un enlace al modelo de usuario, una breve descripción de los intereses del usuario, los documentos seleccionados, presentados de mayor a menor relevancia y para cada documento: título, nombre del autor, nombre de la sección a la que pertenece, fuente, relevancia, resumen automático personalizado, enlace al documento completo en la web del dominio, iconos de realimentación (negativa/positiva).

El resumen de cada documento es extraído automáticamente del texto completo de cada documento aplicando técnicas de selección y extracción de frases. Se utilizan tres heurísticas de selección de frases o sentencias 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 20% sentencias que tengan mayor valor de relevancia. Estas sentencias se concatenarán respetando su orden original para evitar inconsistencias (Kupiec et al., 1995).

La heurística de posición consiste básicamente en dar mayor puntuación a las cinco primeras frases (1, 0.99, 0.98, 0.95, 0.9) de un texto.

La heurística de palabras significativas (Kupiec et al., 1995) consiste en extraer las 8 palabras más significativas de cada texto y comprobar a continuación, cuantas de esas palabras se encuentran en cada frase. De esta forma asignaremos mayor peso a las frases que contengan mayor número de palabras significativas del texto. Para obtener el peso de la frase en el

documento se divide el número de palabras clave del documento que aparecen en la frase por el número total de palabras de la frase.

El objetivo de la heurística de personalización consiste en potenciar aquellas frases que tengan mayor relevancia para un modelo de usuario dado. El cálculo de los pesos para las frases se realiza mediante la combinación lineal ponderada de las similitudes (fórmula de la similitud del coseno del modelo del espacio vectorial) de cada frase con respecto a las palabras clave y los términos de realimentación del modelo de usuario.

Los pesos obtenidos a partir de cada una de las heurísticas son combinados mediante una combinación lineal ponderada para obtener un valor para cada frase para cada una de las noticias con respecto a cada usuario.

#### **4. Metodología de evaluación**

En este apartado se describe la metodología de evaluación utilizada para evaluar el sistema de personalización de contenidos propuesto. En el apartado 4.1 se describe la colección de evaluación generada para poder evaluar los procesos de personalización. En el apartado 4.2 se presentan los tipos de evaluación utilizados para juzgar los procesos de personalización.

##### **4.1. Colección de evaluación**

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.

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 medio de noticias bajadas cada día fue 78.5. Se utilizaron las siguientes 7 secciones: nacional, internacional, deportes, economía, sociedad, cultura y gente. Se El número de usuarios inicial fue 106.

##### **4.2. Tipos de evaluación**

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

La evaluación cuantitativa se basa en la comparación de ranking de noticias con los juicios de relevancia binarios (relevante/no relevante) asociados a los documentos. Esta comparación sugiere la utilización como métricas de recall y precisión normalizados (Salton, 1989). 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, constituye la medida de la efectividad de la personalización realizada.

Por otro lado, la evaluación cualitativa está basada en las respuestas de los usuarios a uno o varios cuestionarios de evaluación donde se le preguntan 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.

## **5. Resultados**

En los apartados 5.1 y 5.2 se describen los resultados obtenidos en la evaluación cuantitativa utilizada en la selección de contenidos y en la presentación de resultados, respectivamente. El apartado 5.3 muestra los resultados de la evaluación cualitativa.

### **5.1. Evaluación de la selección de contenidos**

Para evaluar la selección de contenidos se somete al sistema a las distintas combinaciones de los parámetros asociados a este proceso, es decir, las distintas combinaciones de secciones, categorías, palabras clave y términos de realimentación. Para cada uno de las distintas configuraciones se obtiene un ranking de noticias del cual se obtiene el valor correspondiente de recall y precisión normalizado, para cada usuario, para cada día.

Los resultados obtenidos 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 y recall normalizado. 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 (Díaz&Gervás, 2004).

### **5.2. Evaluación de la generación de resúmenes**

La técnica de evaluación indirecta utilizada 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 (Díaz&Gervás, 2007).

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 a los documentos.

Los resúmenes son de diferentes tipos, dependiendo de la heurística específica que se utilice para generarlos. En este trabajo se han generado 4 tipos de resúmenes: resumen base (20% primeras frases de la noticia completa), resumen genérico (resumen obtenido usando las heurísticas de posición y palabras significativas), resumen personalizado (resumen obtenido usando sólo la heurística de personalización), resumen genérico-personalizado (resumen obtenido utilizando las tres heurísticas).

Los resultados muestran que los resúmenes personalizados ofrecen mejores resultados, con respecto al recall y la precisión normalizada, 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 que los resúmenes personalizados son peores que las noticias completas bajo el mismo criterio.

### **5.3 Evaluación cualitativa**

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.

En términos generales, los usuarios destacan la conveniencia de la combinación de varios sistemas para definir el perfil del usuario. 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 provenientes de combinaciones son significativamente mejores, tanto en selección como en adaptación, que los que utilizan métodos en solitario.

Por otro lado, 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 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.

## **6. Conclusiones**

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. La selección de contenidos a partir de este modelo de usuario ha permitido a los usuarios ajustar mejor sus preferencias sobre los documentos que deseaban recibir. Además la propuesta de personalización presentada ha permitido a los usuarios decidir sobre la relevancia de las noticias recibidas sin necesidad de inspeccionar el texto completo.

La evaluación cuantitativa ha mostrado que la utilización de combinaciones lineales de todos los sistemas de referencia del modelo de usuario ofrece mejores resultados en la selección 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. La evaluación cualitativa ha permitido contrastar los resultados obtenidos en la evaluación cuantitativa a través de las opiniones de los usuarios.

Por tanto, se puede concluir que se ha conseguido el objetivo propuesto inicialmente: disminuir la sobrecarga de información de los usuarios en un sistema de personalización de contenidos Web aplicado a servicios de noticias.

## **Bibliografía**

Billsus, D. & Pazzani, M.J. User Modeling for Adaptive News Access. *User Modeling and User-Adapted Interaction Journal* 10(2-3), (2000), pp. 147-180.

Díaz, A. Integración de técnicas de clasificación de texto y modelado de usuario para la personalización en servicios de noticias. *Tesis doctoral*. Julio 2005.

Díaz, A. & Gervás, P. Adaptive User Modeling for Personalization of Web Contents. *Adaptive Hypermedia and Adaptive Web-Based Systems*. LNCS 3137 (2004), pp. 65-75.

Díaz, A. & Gervás, P. Personalisation in News Delivery Systems: Item Summarization and Multi-Tier Item Selection Using Relevance Feedback. *Web Intelligence and Agent Systems* 3(3), (2005), pp. 135-154.

Díaz, A., Gervás, P. User-model based personalized summarization. *Information Processing and Management* (In Press).

Kupiec, J., Pedersen, O., Chen, F. A trainable document summarizer. *Research and Development in Information Retrieval* (1995), pp. 68–73.

Salton, G. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley Publishing, Reading, Massachusetts, 1989.

Salton G. & McGill, M.J. *Introduction to modern information retrieval*. McGraw-Hill, New York. 1983.

Widyantoro, D.H., Ioerger, T.R., Yen, J. “Learning User Interest Dynamics with a Three-Descriptor Representation”. *Journal of the American Society for Information Science and Technology* 52(3), (2001), pp. 212-225.