

Integración de técnicas de clasificación de texto y modelado de usuario para la personalización en servicios de noticias

Alberto Díaz Esteban

Grupo de Diseño y Optimización de Sistemas Informáticos

ITIS CES Felipe II - UCM

C/ Capitán, 39, Aranjuez, Madrid 28300

adiaz@cesfelipesecondo.com

Resumen: Tesis doctoral en Informática realizada por Alberto Díaz Esteban bajo la dirección de los doctores Pablo Gervás Gómez-Navarro (Univ. Complutense de Madrid) y Manuel de Buenaga Rodríguez (Univ. Europea). El acto de defensa de la tesis tuvo lugar el 13 de Julio de 2005 ante el tribunal formado por los doctores Carmen Fernández Chamizo (Univ. Complutense de Madrid), Manuel Palomar Sanz (Univ. de Alicante), Juan Pavón Mestras (Univ. Complutense de Madrid), Julio Gonzalo Arroyo (UNED), David Bueno Vallejo (Univ. de Málaga). La calificación obtenida fue *Sobresaliente Cum Laude* por unanimidad.

Palabras clave: Personalización de contenidos Web, Modelo de usuario, Evaluación

Abstract: PhD Thesis in Computer Science written by Alberto Díaz Esteban under the supervision of Dr. Pablo Gervás Gómez-Navarro (Univ. Complutense de Madrid) and Dr. Manuel de Buenaga Rodríguez (Univ. Europea). The author was examined in July 13th 2005 by the committee formed by Dr. Carmen Fernández Chamizo (Univ. Complutense de Madrid), Dr. Manuel Palomar Sanz (Univ. de Alicante), Dr. Juan Pavón Mestras (Univ. Complutense de Madrid), Dr. Julio Gonzalo Arroyo (UNED), Dr. David Bueno Vallejo (Univ. de Málaga). The grade obtained was *Sobresaliente Cum Laude*.

Keywords: Web contents personalization, User model, Evaluation

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 a cada tipo de usuario.

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 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.

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 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. 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.

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.

2 Estructura de la tesis

En el capítulo inicial se presentan los objetivos perseguidos y la motivación para el desarrollo de la tesis.

En el Capítulo 2 se revisan los conceptos fundamentales de los sistemas de personalización de contenidos, primero se muestran las distintas formas de representar contenidos, después distintas formas de representar modelos de usuario y posteriormente se divide el proceso de personalización en las 3 funcionalidades principales: selección de contenidos, adaptación del modelo de usuario y presentación de resultados, y para cada una de ellas se presentan tanto las técnicas disponibles como los métodos de evaluación utilizados para juzgar su efectividad.

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

El Capítulo 4 detalla 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.

En el Capítulo 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 y se proponen algunas líneas de trabajo futuro.