Proyecto Mercurio: un servicio personalizado de noticias basado en técnicas de clasificación de texto y modelado de usuario

Alberto Díaz[1], Pablo Gervás[1], José María Gómez[1], Antonio García[2], Manuel de Buenaga[1], Inmaculada Chacón[2], Beatriz San Miguel[1], Raúl Murciano, Enrique Puertas, Matías Alcojor, Ignacio Acero

[1] Departamento de Inteligencia Artificial, Escuela Superior de Informática, Universidad Europea-CEES
Villaviciosa de Odón, Madrid
{alberto, pg2, jmgomez, buenaga, sanmi}@dinar.esi.uem.es
[2] Departamento de Periodismo Especializado, Facultad de Ciencias de la Información, Universidad Europea-CEES
Villaviciosa de Odón, Madrid
{antonio.garcia, inmaculada.chacon}@fcp.cin.uem.es

Resumen. El sistema Mercurio es un servidor personalizado de noticias que trabaja con una representación del cliente basada en los últimos avances sobre modelado de usuario. El servidor de noticias está desarrollado como una aplicación Java que recibe suscripciones de los clientes a través de una página web. Durante el proceso de suscripción el cliente especifica sus preferencias a la hora de recibir noticias, y con ellas se genera un modelo de usuario que se utilizará para enviarle las noticias que puedan interesarle con la frecuencia que haya especificado. El servidor de noticias coopera también con un buscador que permite a los clientes realizar búsquedas puntuales en las noticias del día.

1. Introducción

Los servicios personalizados de información están empezando a ser muy populares en la Web. Muchos de los más importantes periódicos digitales ofrecen a sus usuarios la posibilidad de recibir por correo electrónico una selección de las noticias del día. En todos los servicios revisados por nuestro equipo de investigación, la selección se realiza mediante métodos simples: los usuarios eligen las secciones del periódico en las que están interesados o también ciertas palabras que desean que aparezcan en las noticias que se seleccionen para ellos.

El Proyecto Mercurio ha desarrollado un sistema que aplica técnicas de procesamiento del lenguaje natural y modelado de usuario para seleccionar la información relevante para un determinado usuario. El prototipo del sistema permite a los lectores del periódico recibir periódicamente un mensaje de correo electrónico con las noticias que el sistema encuentra relevantes para los intereses del usuario, previamente definidos en el momento de registrarse en el sistema.

2. Descripción del sistema

Un usuario se conecta al servidor de información y se registra en el sistema. Durante el registro se solicitan una serie de datos esenciales (dirección de correo electrónico, login y password). Después se construye un perfil (o modelo) para cada usuario, que contiene las siguientes informaciones: días de la semana que desea recibir noticias, número máximo de noticias por mensaje e intereses del usuario.

El sistema aplica este modelo de usuario a las noticias del día, usando técnicas de clasificación de texto [2, 3]. Se calcula las relevancia asociada a cada noticia y se construye un ranking. Las noticias con mayor relevancia son enviadas al usuario.

Como servicio adicional, los usuarios pueden realizar búsquedas en las noticias del día.

3. Componentes del sistema

El sistema Mercurio se organiza de acuerdo a las tareas que realiza:

- darse de alta o de baja del servicio, o editar el perfil de usuario.
- recibir un mensaje personalizado todos los días seleccionados por el usuario, conteniendo las noticias más interesantes para él.
- buscar en las noticias del día

El componente de modelo de usuario gestiona los perfiles de usuario y cómo están organizados. Un usuario está representado por un conjunto de informaciones que le caracterizan:

- información general (nombre, login, password, dirección de correo electrónico).

- información sobre sus preferencias (días de la semana que desea recibir mensajes, máximo número de noticias por mensaje, desactivación temporal del servicio).
- información sobre los intereses del usuario: (1) secciones, (2) categorías generales, (3) términos.

Las categorías generales corresponden a las categorías de primer nivel de Yahoo! Spain y constituyen un sistema alternativo de clasificación para los usuarios.

El modelo de usuario puede ser cambiado por el usuario todas las veces que considere necesario teniendo en cuenta que el sistema se ejecuta todos los días una sola vez a primera hora de la mañana, en el momento en que las noticias están disponibles.

El componente de búsqueda permite realizar una búsqueda básica sobre las noticias del día, por si el usuario está interesado en alguna noticia de manera puntual y no quiere reflejarlo en su perfil.

La componente de envío de mensajes consulta los perfiles de usuario cada día para construir un mensaje para todos los usuarios con las noticias del periódico más interesantes del día. El mensaje que recibe el usuario contiene: el nombre del usuario, la fecha y para cada noticia, su título, su relevancia, un resumen y un enlace a la noticia completa en el periódico digital. Adicionalmente, al final del mensaje se muestran al usuario los intereses que tiene seleccionados en su perfil: secciones, categorías generales y términos.

3. Filtrado de información usando tareas de clasificación de texto

La representación de las noticias se obtiene aplicando el modelo del espacio vectorial [5] a su texto. La representación de las categorías generales se realiza de igual forma con el texto asociado a las páginas web de las categorías de primer nivel de Yahoo! Spain. Los términos se representan usando este modelo.

En el proceso de selección se aplica categorización de texto [4] con las categorías respecto de las noticias y recuperación de información [4] con los términos con respecto a las noticias. Adicionalmente las noticias son procesadas para comprobar si pertenecen a alguna de las secciones seleccionadas por el usuario. Los diferentes resultados obtenidos se integran usando el nivel de interés asignado por el usuario a los diferentes sistemas de referencia.

4. Conclusiones

El sistema Mercurio implementa un servicio de envío personalizado de noticias basado en un modelo de usuario que captura los diferentes intereses del usuario. Con este modelo y técnicas de clasificación de texto se filtran las noticias relevantes para un determinado usuario. El sistema propuesto tiene todos los aspectos deseables para estos sistemas: uso de más información que secciones y términos para describir los intereses del usuario, efectividad, personalización, facilidad de uso y ausencia de ruido. The results of the evaluation [1] show the general satisfaction of users with the system.

Bibliografía

[1] Díaz, A., Gervás, P., García, A., Evaluating a User-Model Based Personalization Architecture for Digital News Services, To appear in *Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libraries (ECDL'2000)*, Lecture Notes in Computer Science, Springer-Verlag.

[2] Díaz, A., Buenaga, M., Ureña, L. A., García, M. (1998) ‘Integrating Linguistic Resources in an Uniform Way for Text Classification Tasks’, First International Conference on Language Resources & Evaluation, Granada (Spain), 1998.

[3] Gómez Hidalgo, J. M. and Buenaga, M. (1997) 'Integrating a Lexical Database and a Training Collection for Text Categorization', ACL/EACL Workshop on Automatic Information Extraction and Bulding of Lexical Semantic Resources for NLP, Madrid (Spain), 1997.

[4] Lewis, D. (1992). Representation and learning in information retrieval. Ph.D. Thesis, Department of Computer and Information Science, University of Massachusetts. 1992.

[5] Salton G. & McGill, M.J. (1983). *Introduction to modern information retrieval.* McGraw-Hill.1983.