Proyecto Hermes: Servicios de Personalización Inteligente de Noticias mediante la Integración de Técnicas de Análisis Automático del Contenido Textual y Modelado de Usuario con Capacidades Bilingües[?]

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

[1] Departamento de Inteligencia Artificial, Escuela Superior de Informática
{alberto, buenaga, giraldez, jmgomez, sanmi, epuertas, murciano}@dinar.esi.uem.es
nachoa@telcom.es, alcojor@retemail.es
[2] Departamento de Periodismo Especializado, Facultad de Ciencias de la Información
{antonio.garcia, inmaculada.chacon}@fcp.cin.uem.es
Universidad Europea-CEES, Villaviciosa de Odón
28670 Madrid, España

Pablo Gervás

Departamento de Sistemas Informáticos y Programación, Universidad Complutense de Madrid, Ciudad Universitaria, 28040 Madrid, España

pgervas@sip.ucm.es

Resumen El proyecto Hermes tiene como objetivo el desarrollo de un sistema personalizado inteligente de acceso a la información en un entorno bilingüe, español e inglés. El sistema proporciona una alta efectividad e información especialmente adaptada al cliente, basándose en la utilización de técnicas avanzadas del contenido textual y modelado de usuario. Un objetivo principal del proyecto Hermes radica en la extensión de las tecnologías vigentes para entornos monolingües al campo bilingüe. 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.

1. Introducción

Actualmente hay más de 6000 periódicos digitales en Internet. La mayoría de ellos suministran información redundante, muchas veces sin verificar o de manera incompleta. Además una persona no espera que un periódico electrónico sea simplemente otra versión del periódico tradicional: la personalización es un valor añadido que se está ofreciendo a los lectores de los periódicos digitales. Muchos de los más importantes servicios de noticias ofrecen a sus usuarios la posibilidad de recibir por correo electrónico una selección de las noticias del día. Sin embargo, en la mayoría de ellos 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.

Una definición tan simple de los intereses de un usuario puede hacer que mucha de la información recibida sea realmente irrelevante. La creación de modelos de usuario que mejoren la definición de esos intereses junto con la integración de tareas de clasificación de texto permiten mejorar la selección de las noticias relevantes para cada usuario [1].

Además, cada vez se hace más patente la necesidad de manejar informaciones en diferentes idiomas. En el sistema Hermes se ha desarrollado una extensión de las tecnologías vigentes para entornos monolingües al campo bilingüe para seleccionar la información relevante para un determinado usuario. El prototipo del sistema permite a sus usuarios recibir periódicamente un mensaje de correo electrónico con las noticias, de dos periódicos digitales en dos idiomas diferentes, que el sistema encuentra relevantes para los intereses del usuario.

[?] Este proyecto ha sido financiado parcialmente por el Ministerio de Ciencia y Tecnología (PROFIT, 2000/020)

2. *Descripción del sistema*

El usuario nada más entrar en la página web del sistema Hermes, puede elegir entre dos posibilidades (español e inglés). El idioma seleccionado será el que utilizará el sistema para manejar el modelo de ese usuario, es decir, cada usuario confecciona su perfil en un único idioma, el que él seleccione.

En la página de alta el usuario introduce sus datos personales (nombre, dirección de correo electrónico, número máximo de noticias por mensaje, etc.) y sus intereses. Dentro de los intereses hay una parte basada en la estructura de las noticias y otra parte basada en el contenido. Este perfil representa los intereses a largo plazo del usuario.

La parte basada en la estructura está formada por las secciones de los periódicos. El usuario puede seleccionar las secciones del periódico en español y las secciones del periódico en inglés en las que está interesado.

La parte basada en el contenido está formada por un conjunto de categorías generales y un conjunto de términos. Las categorías son las correspondientes al primer nivel de Yahoo y son independientes del idioma. Los términos están en el idioma seleccionado por el usuario.

El sistema construye para cada usuario dos modelos, uno en cada idioma y aplica cada modelo a las noticias en el mismo idioma.

La única información que hay que traducir de un idioma a otro es la información relativa a los términos, para ello, se realiza una desambiguación de cada término y posteriormente se traduce.

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

El mensaje que recibe el usuario contiene: el nombre del usuario, la fecha y para cada noticia, su título, su relevancia, la sección a la que pertenece, un resumen adaptado al usuario y un enlace a la noticia completa en el periódico digital. Además, se muestran al usuario los intereses que tiene seleccionados en su perfil: secciones, categorías y términos.

Adicionalmente el usuario puede realimentar al sistema votando afirmativa o negativamente sobre las noticias que recibe. El título y el resumen de la noticia realimentada son utilizados para obtener otra información sobre los intereses del usuario: los términos de realimentación. Estos términos reflejan los intereses a corto plazo del usuario y son utilizados como otro interés en el siguiente envío de noticias. Con estos términos hay que realizar el mismo proceso de traducción que con los términos a largo plazo.

3. *Filtrado de información usando tareas de clasificación de texto.*

La representación de las noticias se obtiene aplicando el modelo del espacio vectorial a su texto. La representación de las categorías generales se realiza aplicando técnicas de aprendizaje automático al texto asociado a las páginas web de las categorías de primer nivel de Yahoo en cada idioma. Los términos se representan usando el mismo modelo.

En el proceso de selección se aplica categorización de texto con las categorías respecto de las noticias y recuperación de información con los términos con respecto a las noticias, aplicando el modelo en el mismo idioma que la noticia. 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 Hermes implementa un servicio de envío personalizado de noticias basado en un modelo de usuario que captura intereses a largo y corto plazo del usuario. La creación de un modelo por idioma y la utilización de técnicas de clasificación de texto permiten realizar el filtrado de las noticias relevantes para un determinado usuario.

Bibliografía

[1] Díaz, A., Gervás, P., García, A. (2000). Evaluating a User-Model Based Personalisation Architecture for Digital News Services. Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries (ECDL).