Procesamiento del Lenguaje Natural, num. 29 (2002), pp.289-290 recibido 12-07-2002; aceptado 15-07-2002

OmniPaper: Acceso Inteligente a Periódicos Europeos

José Carlos González, Julio Villena
DAEDALUS – Data, Decisions and Language, S.A.
{jgonzalez, jvillena}
@daedalus.es

Francisco Bueno, Ana M. García Serrano, Alberto Ruiz Cristina
Dep. de Inteligencia Artificial
Facultad de Informática
Universidad Politécnica de Madrid
{bueno, agarcia, aruiz}
@dia.fi.upm.es

Paloma Martínez Fernández
Grupo de Bases de Datos Avanzadas
Departamento de Informática
Universidad Carlos III de Madrid
pmf@inf.uc3m.es

Resumen: El proyecto OmniPaper (proyecto IST-2001-32174, www.omnipaper.org) pretende ofrecer un medio de acceso personalizado y unificado a las noticias de periódicos europeos. Para facilitar el acceso a los artículos se combina la navegación por una taxonomía multilingüe de temas con búsquedas a texto completo y otras técnicas que incluyen el manejo de metadatos.
Palabras clave: recuperación de información, consulta multilingüe

Abstract: The aim of the OmniPaper project (IST-2001-32174, www.omnipaper.org) is to offer a personalized and standarized interface to the articles of European newspapers. In order to improve this access, the browsing through a multilingual taxonomy of subjects is combined with full-text searching and other techniques which involve metadata handling.
Keywords: information retrieval, multilingual query

1 Consorcio

- Katholieke Universiteit Leuven - Research Group on Document Architectures (project co-ordination)
- Universidad Politécnica de Madrid - Dep. de Inteligencia Artificial (Madrid)
- DAEDALUS - Data, Decisions and Language, S.A. (Madrid)
- Universidade do Minho - Departamento de Sistemas de Informação (Oporto)
- Center for Usability Research and Engineering (Vienna)
- My News (Barcelona)
- Pressetext.austria Nachrichtenagentur (Vienna)
- Mediargus (Brussels)

2 Descripción del proyecto

Desde que comenzó el proceso de digitalización de los grandes recursos de información, el problema de la recuperación eficiente de conocimiento ha cobrado más importancia, ya que los recursos digitales no pueden ser realmente útiles sin un sistema de recuperación adecuado. Ante esta situación, los editores de periódicos han invertido un esfuerzo considerable en estructurar sus contenidos de forma que se facilite su recuperación; sin embargo, la heterogeneidad en los formatos de almacenamiento de noticias y en la metainformación que las describe ha hecho imposible definir un recurso accesible de forma unificada. Por tanto, la búsqueda de noticias aún se hace en la mayoría de los casos mediante robots de búsqueda en texto completo, de forma que la calidad del resultado depende en gran medida de la sofisticación de la consulta introducida por el usuario.

OmniPaper se enmarca dentro de los proyectos de recuperación inteligente de información. En él se investigan técnicas para alcanzar una nueva forma de acceder a las noticias; algunas de ellas están relacionadas con la Inteligencia Artificial, como las técnicas de minería de datos para la extracción de rasgos de usuarios o de términos relevantes en los documentos.

3 Objetivos principales del proyecto

El objetivo clave del proyecto OmniPaper es la creación de una capa de navegación y enlace situada por encima de recursos de información distribuidos, en un entorno con capacidad de

ISSN 1135-5948

aprendizaje. Como resultado final, OmniPaper ofrecerá un medio de acceso inteligente y uniforme a los periódicos europeos en internet; el usuario no necesitará buscar direcciones específicas de periódicos, ni adaptarse a las diferencias entre sus interfaces de usuario y métodos de búsqueda. Esta nueva puerta de entrada le permitirá buscar noticias en todas las bases de datos con una sola consulta, expresada en un solo idioma. Independientemente de la lengua en que se encuentren los diferentes archivos consultados, el usuario recibirá las noticias en su idioma.

4 Propuesta de arquitectura del proyecto

La arquitectura de OmniPaper parte de un conjunto de archivos de noticias distribuidos, cada uno de ellos con sus propios entornos de ejecución, formatos de representación y mecanismos de acceso e indexación. El protocolo SOAP (*Simple Object Access Protocol*) se utiliza para garantizar un acceso uniforme a dichos archivos. En cuanto a la búsqueda inteligente, se apoya en la riqueza de la indexación y en estructuras de metainformación (RDF y Topic Maps).

A nivel de organización, OmniPaper se divide en varios módulos. Se distingue el nivel local (recuperación de información distribuida) del nivel de conocimiento global. El nivel local se construye sobre cada uno de los proveedores; para ello es necesario estudiar y probar técnicas de recuperación de información a partir de fuentes distribuidas.

Por su parte, el nivel de conocimiento es una taxonomía que almacena los conceptos extraídos de los diferentes archivos, así como distintas relaciones semánticas entre ellos y sus términos asociados, en diferentes idiomas. Cada concepto está relacionado mediante índices con los documentos relacionados, con independencia del archivo del que procedan: el nivel de conocimiento global unificará los niveles locales de forma estructurada. De esta forma será posible la navegación a través de archivos en un entorno multilingüe, dando como resultado un nivel de conocimiento que engloba a todos los archivos. Por encima del nivel global, la interfaz de usuario garantizará una presentación sencilla e interactiva del nivel de conocimiento.

A partir de esta taxonomía surge el concepto de *consulta guiada*, que puede ser modificada dinámicamente por el usuario: si los términos de la consulta son demasiado generales, se le ofrece una lista de opciones para enfocarla mejor; si por el contrario son demasiado específicos, la consulta guiada ofrecerá sinónimos y temas relacionados, de forma que el usuario podrá obtener resultados sobre temas en los que estaba interesado aunque no lo hubiera expresado en la consulta.

La información sobre las elecciones del usuario en la navegación será utilizada para aprendizaje del sistema. De esta forma se garantiza que la taxonomía contendrá enlaces actualizados, relevantes y útiles.

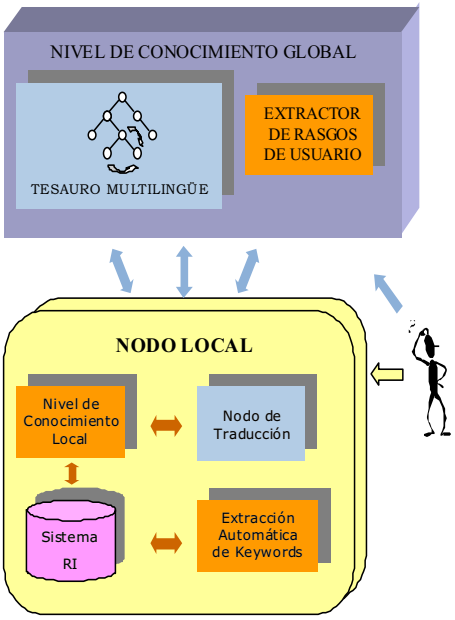

Figura 1: Arquitectura del Sistema

La Figura 1 muestra la aproximación *bottom-up* que se utilizará en el proyecto. El nivel local se construye aumentando las capacidades de cada proveedor de contenidos de forma individual con técnicas de extracción de palabras clave, tratamiento de consultas de usuario y traducción de términos; este trabajo de mejora variará en función de las capacidades actuales de cada proveedor. El nivel de conocimiento global proporciona un interfaz común de acceso a todos los niveles locales, añadiendo además el tratamiento de los diferentes idiomas y la adaptación del sistema a las preferencias del usuario.