Análisis y expansión de consultas en lenguaje natural para mejora de la búsqueda en Web

Alberto Ruiz[1], Paloma Martínez[2] y Ana García-Serrano[1]

[1]Grupo ISYS-Hermeneumática
Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid
{agarcia,aruiz}@isys.dia.fi.upm.es

[2] Grupo de Bases de Datos Avanzadas
Departamento de Informática
Universidad Carlos III de Madrid
pmf@inf.uc3m.es

1. Resumen

El proceso tradicional de búsqueda en Web se encuentra limitado por los lenguajes de consulta y por la carencia de información semántica sobre el dominio al que se refiere el usuario. Esto provoca que el sistema no recupere todos los resultados relevantes y sí obtenga, por el contrario, resultados que nada tienen que ver con la consulta original. El proyecto MESIA (CAM 07T/0017/1998) intenta paliar esta situación en el servidor Web de la Comunidad Autónoma de Madrid (www.comadrid.es), actuando como interfaz entre el usuario y el buscador Altvista.

2. Objetivos

El sistema MESIA facilita al usuario la comunicación con el motor de búsqueda, actuando como interfaz en dos niveles: antes de la búsqueda, recibe la consulta del usuario escrita en lenguaje natural y, posteriormente, la convierte en una consulta booleana. Durante este proceso se produce una expansión de la consulta mediante recursos lingüísticos que mejora sensiblemente los resultados.

Después de la búsqueda, MESIA incorpora información sobre el dominio al proceso, permitiendo la expansión semántica de resultados: una vez identificado el tema de la consulta, a los resultados obtenidos se añaden enlaces sobre asuntos relacionados con dicho tema. Además, esta información sirve también para ordenar los resultados según su relevancia para la consulta.

3. Descripción del sistema

La figura 1 describe la arquitectura del prototipo actual de MESIA (http://tornado.dia.fi.upm.es/mesia/mesia_demo.hml), que ha sido implementado en el entorno CIAO-Prolog [1].

A continuación se describen los dos módulos principales: el módulo de expansión de la consulta y el de ampliación y ordenación de resultados.

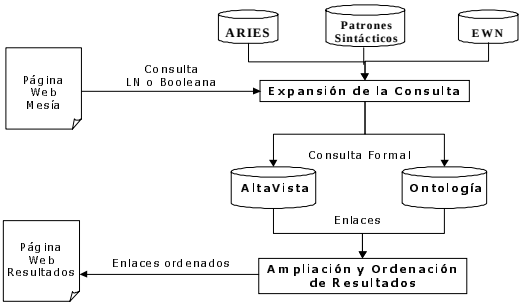

Figura 1: Arquitectura de MESIA

Cuando se recibe la consulta del usuario, en primer lugar se realiza un análisis morfológico de las palabras que la forman, para poder etiquetar cada palabra con su categoría (*part of speech*).

Para realizar este análisis se utiliza la plataforma léxica ARIES [2]. Un segmentador de frases, [3], descompone la consulta en sintagmas nominales, preposicionales y verbales con el fin de resolver la ambigüedad del proceso de etiquetado morfológico. Además, este análisis sintáctico parcial permite extraer los términos relevantes de la consulta (núcleos y modificadores, en una primera aproximación).

Estos términos son expandidos a continuación mediante EWN para castellano, [4], una base de datos semántica que permite obtener sinónimos de cada palabra. El siguiente paso es obtener las variantes morfológicas (género y número) para todos los términos; en caso de ser verbos, sólo se mantiene el infinitivo. Por último, se construye la consulta booleana en forma normal conjuntiva.

El resultado de este proceso se envía al modo de Búsqueda Avanzada de Altavista. Adicionalmente, al campo "*Ordenar Por*" de dicho motor de búsqueda se envían las variantes morfológicas de la consulta origina; esto permite dar prioridad a los términos originales de la consulta para mejorar la clasificación de los documentos relevantes. Además, se ha

llevado a cabo una evaluación con consultas de usuarios para estudiar la influencia de las expansiones realizadas a los términos relevantes de las consultas

En cuanto a la ampliación y ordenación de resultados, cuando un usuario hace una consulta, probablemente también estará interesado en temas relacionados con la misma; estos otros temas no están necesariamente descritos por las mismas palabras, de forma que quizá no aparezcan como resultado del proceso de búsqueda tradicional, incluso después de la expansión lingüística.

MESIA incorpora información semántica al proceso de búsqueda a través de una ontología implementada en XML que almacena información sobre el dominio; dicha ontología consiste en un árbol de conceptos relacionados en que cada nodo representa uno de los temas que pueden encontrarse en el dominio. Los nodos se identifican por un conjunto de palabras clave que lo describen, e incluye un conjunto de enlaces relacionados con el tema. La figura 2 muestra algunos nodos de la ontología de dominio de MESIA, con sus palabras clave.

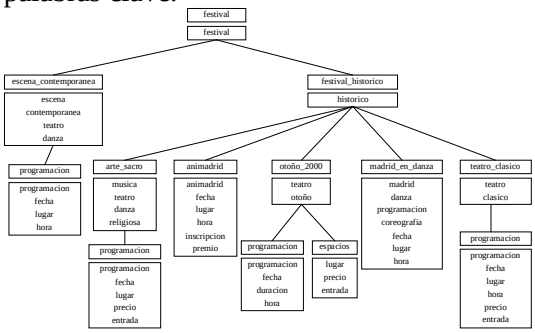

Figura 2: Ontología de dominio

La integración de la ontología en el proceso de búsqueda se realiza de la siguiente forma: como se aprecia en la figura 1, la consulta expandida es enviada simultáneamente al motor de búsqueda y a la ontología. Ambos recursos obtienen sus resultados y, posteriormente, el módulo de ampliación los junta y ordena. Al funcionar por separado, si alguno de los dos recursos no estuviese disponible por cualquier motivo, el otro se mantendrá operativo y el usuario obtendrá resultados.

Para extraer información de la ontología se realiza un acceso basado en un sistema de pesos, que se puede describir brevemente de la siguiente forma: Las palabras de la consulta expandida se comparan con todas las palabras clave de todos los nodos de la ontología, de forma que cuando hay coincidencia, el nodo en que estaba la palabra incrementa su peso (valor numérico). Este peso se propaga, con menor valor, a los nodos situados cerca de dicho nodo en la estructura de la ontología.

Al finalizar el proceso, los nodos son ordenados por el peso obtenido (que indica relevancia), y sus enlaces asociados se presentan agrupados bajo el título del nodo; esto supone una considerable mejora en la presentación de los resultados.

La ontología requiere un importante trabajo de actualización. Por el momento se supone que este trabajo se realiza manualmente; en el futuro se abordará la clasificación automática de páginas en la ontología.

Figura 3: Interfaz MESIA

Referencias

1. F. Bueno, D. Cabeza, M. Carro, M. Hermenegildo, P. López, and G. Puebla, The Ciao Prolog System: A Next Generation Logic Programming Environment, TR CLIP 3/97.1(www.clip.dia.fi.upm.es/Software/ Ciao/), 1999.

2. J. M. Goñi, J. C. González and A. Moreno, ARIES: A lexical platform for engineering Spanish processing tools. Natural Language Engineering, vol 3 no 4, pp. 317-345, 1997.

3. P. Martínez and A. García-Serrano, The role of knowledge-based technology in language applications development. Expert Systems with Applications, vol. 19, no 2, pp. 155-160, 2000.

4. P. Vossen, The EuroWordNet Base Concepts and Top Ontology. Version 2. EuroWordNet (LE 4003) Deliverable, 1998.