Utilizando recursos lingüísticos para mejora de la recuperación de información en la Web

Paloma Martínez Fernández†, Ana García Serrano‡

†Universidad Carlos III de Madrid, Departamento de Informática
Grupo de Bases de Datos Avanzadas,
Avda. de la Universidad 30, 28911 Leganés, Madrid, España
pmf@inf.uc3m.es

‡Universidad Politécnica de Madrid, Departamento de Inteligencia Artificial,
Grupo ISYS,
Campus de Montegancedo S/N, 28660, Madrid, España
agarcia@dia.fi.upm.es

Resumen

Una de las hipótesis para la mejora de la interacción persona-ordenador se basa en el uso del lenguaje natural; en un primer nivel básico se emplea conocimiento lingüístico simple en aquellos procesos no muy complejos involucrados en la interacción. En este trabajo se muestran aspectos relacionados con la integración de la tecnología disponible de tratamiento de lenguaje natural en el desarrollo de un metabuscador que alcance un mayor grado de acierto en la recuperación de información realizada por un buscador tradicional así como en el tratamiento posterior de los documentos recuperados. En particular, se describe el proceso realizado para la extensión de las consultas de los usuarios con información lingüística empleando dos recursos léxicos para el castellano: ARIES para el tratamiento de la morfología y EuroWordnet para el tratamiento de la semántica.
Este trabajo forma parte del sistema MESIA, Modelo computacional para extracción selectiva de información de textos cortos, que amplía la búsqueda habitual (consulta y presentación de resultados) con nuevas capacidades morfológicas y semánticas y analiza otros aspectos obtenidos a partir de la estructura de las páginas, del tratamiento lingüístico de algunas de las unidades de texto seleccionadas automáticamente y de la experiencia de uso.
El sistema está diseñado para el sitio Web de la Comunidad Autónoma de Madrid (CAM), lo que representa una restricción en la cantidad de información disponible, pero se mantiene la problemática general de la búsqueda de información ya que la información contenida en estas páginas abarca prácticamente todas las categorías informativas que la Administración puede ofrecer al ciudadano.

Palabras clave: Interacción Web, consultas en lenguaje natural, recursos lingüísticos.

1. Introducción

En la segunda mitad de la década de los noventa, la implantación en el ámbito social de Internet ha motivado una creciente demanda de nuevas formas de gestión y métodos eficaces para la búsqueda de información en la Web, lo que genera necesidades de acceso a la información con mayores prestaciones que las que proporcionan las soluciones actuales. Si se trata de las páginas de un organismo oficial, la precisión en la oferta de información al ciudadano es uno de los objetivos críticos para este tipo de sistemas. La información pertinente y la búsqueda eficiente (de acuerdo con criterios variables) son dos aspectos cruciales para la aceptación por parte del usuario, de cualquier sistema de información a través de Internet que utilice una organización para su publicidad, negocio o para cumplir alguno de sus objetivos.

En esta presentación nos centraremos en los aspectos relacionados con el resultado de la búsqueda a partir de una consulta en lenguaje natural.

Actualmente los buscadores existentes (AltaVista, Yahoo, y otros) se basan en análisis estadísticos que aportan una forma de discriminación y selección de páginas Web relacionadas con una consulta; el crecimiento exponencial de la información en Internet hace que la simple equiparación de cadenas (búsqueda basada en palabras clave) produzca muchos más documentos de los necesarios (información irrelevante mezclada con la relevante) y que documentos que deberían aparecer como resultado de una búsqueda no lo hagan por no contener explícitamente los términos de la consulta sino otras palabras o expresiones relacionadas con la consulta, Baeza-Yates y Ribeiro-Neto(1999).

Basándose en la hipótesis de que para mejorar los resultados de la búsqueda es necesario modificar automáticamente la consulta, se pueden plantear estrategias que tengan en cuenta otros aspectos relacionados con los documentos para su selección, así como proporcionar mecanismos que incorporen criterios de búsqueda para mejorar la especificación de petición (a través de la experiencia adquirida).

Los trabajos realizados hasta el momento se concretan en:

La modificación de la consulta original. El sistema transforma la consulta realizada en un lenguaje cercano al natural en una consulta formal, extrayendo los términos significativos y extendiéndolos mediante la inclusión de variantes morfológicas o sinónimos. El resultado de este proceso se guarda en una estructura que contiene información sobre la consulta original y las efectuadas por el buscador.

La clasificación de los documentos que forman el resultado de la búsqueda. El propio sistema de búsqueda extraerá del documento la información necesaria para su identificación a partir de una consulta (*content*). Esta información permitirá aplicar diferentes criterios obtenidos a partir del análisis del dominio o de forma experimental, ante la falta de una estructura unificadora que permita clasificar los documentos con un grado de certeza absoluto. Actualmente, estos criterios son de dos tipos, estructurales (los documentos pueden admitir cuatro formas posibles de acuerdo a la forma de su contenido) y semánticos (temática, objetivo divulgativo del texto, etc.). El resultado de este proceso se refleja en una estructura de rasgos que se genera para cada documento analizado y que se utiliza en el siguiente paso.

La acumulación de la experiencia. El sistema dispone de un gestor tanto del conocimiento extraído de los documentos como de los propios documentos (denominado bibliotecario). Está prevista la utilización de perfiles de usuario que permitirán decidir si la consulta se envía al gestor de conocimiento (que incorpora conocimiento sobre las consultas mas frecuentes de cada tipo de usuario, o el resultado de análisis de los documentos correspondientes a consultas ya realizadas) o se lanza una nueva búsqueda. El modelo de usuario disponible actualmente es muy simple (una ontología con una descripción basada en el uso previsto del sistema para cada tipo de usuario) pero permite incorporar en algún caso condiciones a la consulta formal para intentar delimitar la respuesta del metabuscador.

El tratamiento automático tanto de la consulta como de los textos significativos incluidos en los documentos para la generación automática de la estructura semántica, exige la articulación tanto de conocimientos lingüísticos (generales y terminológicos) como del dominio y de control del proceso. Para el diseño basado en el conocimiento del sistema para búsquedas selectivas en castellano se han identificado los siguientes tipos de conocimiento a partir del de análisis manual realizado y uso previsto del sistema:

1. Conocimiento sobre la estructura de los documentos para su clasificación de acuerdo con diferentes criterios.
2. Conocimiento lingüístico sobre el sublenguaje del domino (o dominios), el léxico terminológico y del análisis guiado por

expectativas basado en expresiones significativas.

3. Conocimiento sobre los usuarios que realizan las consultas: preferencias (de acuerdo con el histórico) y otras restricciones (positivas o negativas).

Para que este conocimiento sea operativo es necesario desarrollar un sistema software que incorpore y articule convenientemente cada tipo de conocimiento. Es el conocimiento lingüístico del que nos ocuparemos mas extensamente en este artículo, ya que es el requerido para realizar el proceso automático del lenguaje natural (LN) de las consultas de los usuarios, y que incorpora recursos ya disponibles como son los léxicos terminológicos y generales, o etiquetadores morfológicos.

2. Conocimiento lingüístico necesario para la recuperación de información

En los últimos años, se han incorporado algunas técnicas de análisis del lenguaje natural a los buscadores existentes Liddy (1998) dado que los métodos puramente estadísticos no llegan a alcanzar los resultados deseados. A continuación se indican para los distintos niveles de análisis lingüístico, los recursos y técnicas lingüísticas que pueden incorporarse para este fin:

1. Nivel morfológico: Es el nivel lingüístico que mas aparece en los sistemas de recuperación de información (RI). Así, los algoritmos de stemming (extracción de raíces) pueden ayudar a evitar que documentos relevantes a una consulta no sean eliminados (por ejemplo, si no se hace este análisis para las formas plurales de los nombres, entonces los documentos que incluyan las formas singulares no serán recuperados). Hay que destacar que el procesamiento morfológico no ofrece las mismas posibilidades para todos los idiomas. En el caso del castellano, al ser una lengua altamente flexiva, es decir con una rica morfología, este análisis puede proporcionarnos mucha información en el proceso de RI, pero plantea problemas de eficiencia.

2. Nivel léxico: Este nivel lingüístico puede emplearse en RI tanto para el etiquetado morfosintáctico ad-hoc de los términos de las consultas (categorías gramaticales, nombres, verbos, adjetivos, etc.) como en la utilización de léxicos en los que se encuentran los rasgos gramaticales y semánticos de las palabras. Este nivel se evidencia en el conocimiento contenido en los thesauros y otros recursos similares. Así, nos pueden interesar las relaciones sintagmáticas y paradigmáticas de los términos que pueden ayudarnos en la formulación (semi)automática de las consultas (por ejemplo, rasgos especiales de las palabras como nombres propios, acrónimos, etc.).

3. Nivel sintáctico: A partir de la salida del etiquetado morfosintáctico y con el fin de detectar frases o grupos significativos para un dominio o sublenguaje concreto, este conocimiento se puede aplicar tanto a de la consulta como al texto de un documento. De esta forma se pueden conseguir mejores términos de indexación que representan el contenido de los documentos así como palabras clave mas precisas en las consultas (no es lo mismo buscar documentos sobre “conciertos” que sobre “conciertos de jazz ”).

4. Nivel semántico: Este nivel se refiere a la interpretación del significado de las oraciones como una unidad, en contraposición con el significado individual de las palabras o sintagmas que las componen. Algunos de los fenómenos lingüísticos que se pueden tratar conciernen a la desambiguación semántica, a la identificación de las relaciones del verbo y sus argumentos en una oración o a la expansión de una consulta mediante la adición de todos los sinónimos equivalentes de los términos de la consulta.

5. La expansión de términos puede realizarse empleando fuentes léxicas como EuroWordNet (EWN), Vossen (1997), Gonzalo et al. (1998), o un thesauro; sin embargo, el reto consiste en añadir sólo aquellos términos que son pertinentes y que suponen la expansión acertada del significado particular de la palabra en la consulta.

6. Otra aproximación del procesamiento semántico es la producción de vectores semánticos para representar documentos y consultas, Voorhees (1999), pero su efectividad también se basa en que el significado apropiado de cada término se haya determinado previamente a su inclusión en el vector semántico.

7. Nivel de discurso: En este nivel se tratan aquellos aspectos que permiten identificar algunos de los principios de estructura y organización que implícitamente utilizan los autores de los documentos y consultas. El análisis del discurso busca el papel específico que una pieza de información desempeña en un documento, por ejemplo, si es una conclusión, una opinión, un hecho, una predicción, etc. Adicionalmente, el reconocimiento y resolución

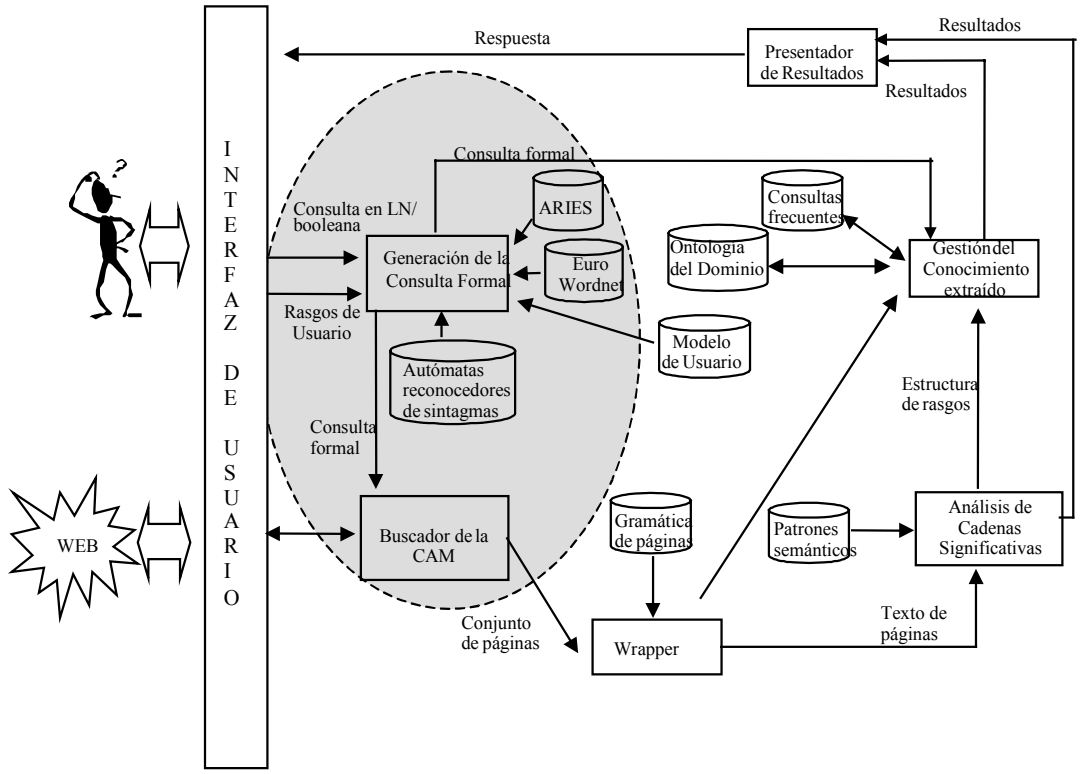

Figura 1: Arquitectura de MESIA

de una anáfora (fenómeno lingüístico de referencia de un elemento de la oración mediante un pronombre, por ejemplo) puede mejorar tanto el análisis de los documentos como la de las consultas, aunque pocos sistemas lo incorporan actualmente por la dificultad intrínseca de la anáfora, Palomar et al. (2001).

8. Nivel pragmático: Este nivel concierne con diferentes aspectos del entorno externo (a la pregunta y al discurso) que influyen en las interacciones entre el usuario y el sistema. Del mismo modo que un buen bibliotecario puede extraer de los usuarios diferente información como por ejemplo la finalidad para la que planean utilizar el documento que están buscando, los sistemas de RI se beneficiarían al conocer las necesidades del usuario en el contexto de su historia y de sus objetivos particulares. Por lo tanto, puede también plantearse la incorporación algunos aspectos básicos de la comunicación en la interfaz de usuario de un sistema RI y que contengan referencias al modelo de usuario.

Existen motores de búsqueda que incorporan alguno de estos niveles de conocimiento lingüístico (morfología, sintaxis, léxico, semántica) sobre todo en prototipos de investigación más que en buscadores comerciales. Particularmente, se hace uso de listas de parada (palabras demasiado frecuentes en el lenguaje y que no aportan información de utilidad como son las preposiciones, adverbios, determinados verbos, etc.), algoritmos de extracción de raíces para el inglés, thesauros (relacionan las palabras de un determinado dominio mediante clases), bases de datos léxicas como Wordnet, Miller (1995), para recuperación de información mediante la equiparación de conceptos en vez de palabras o términos así como para la expansión de las consultas y léxicos morfológicos (que permiten utilizar información morfosintáctica para mejorar la aplicación de los métodos estadísticos).

2.1. Uso de recursos lingüísticos en MESIA

La incorporación de técnicas lingüísticas al buscador se ha abordado en la primera versión de MESIA, a partir de la integración de diferentes recursos lingüísticos disponibles actualmente para el castellano: ARIES y EuroWordnet.

ARIES (http://www.mat.upm.es/~aries/), Goñi et al. (1997), es un léxico morfológico para el castellano desarrollado por la Universidad Politécnica de Madrid y la Universidad Autónoma de Madrid (licencia de uso). Está formado por un léxico español de 38.500 lemas y 600 morfemas flexivos, varias utilidades de acceso y mantenimiento y un analizador/generador morfológico. Para su

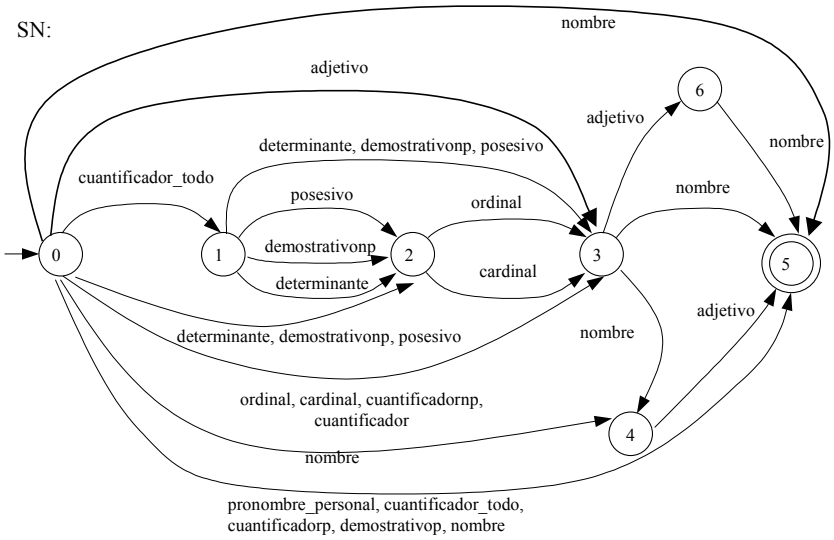

Figura 2: Descripción del autómata reconocedor de Sintagmas Nominales

integración en MESIA ha sido necesario traducir la base de datos léxica a PROLOG, Bueno et al. (1999).

En particular, en el sistema MESIA se utiliza un generador morfológico basado en el formalismo DCG (Definite Clause Grammar) para formación de palabras. Este generador permite, por ejemplo, obtener las formas *doctor*, *doctora*, *doctores* y *doctoras* a partir del lema doctor así como obtener sus categorías morfológicas. La palabra doctor tiene asociada la siguiente entrada en el léxico ARIES:

```
doctor
    cat      = n  /* nombre */
    concat   = wl /* palabra que acepta
               morfema de número */
    agr gen  = masc  /* género masculino */
    agr num  = sing /* número singular */
    nut      = plu2 /* formación del plural */
    lex      = doctor  /* lema */
```

EuroWordNet (http://www.let.uva.nl/~ewn/), Vossen (1997), Gonzalo et al. (1998b), es una base de datos léxica que está estructurada en una ontología construida con varias relaciones semánticas entre palabras (nombres, adjetivos, verbos, etc.) en inglés, español, alemán e italiano.

Tiene dos niveles de estructura, una ontología superior de conceptos que refleja diferentes relaciones explícitas de oposición (v.g. animado e inanimado) y que puede verse como una representación de los distintos campos semánticos del vocabulario de EuroWordNet y una jerarquía de etiquetas de dominio que relacionan conceptos según distintos temas, v.g. deportes, deportes de invierno, deportes de agua, etc. Las relaciones semánticas más importantes de EuroWordNet son la sinonimia, antonimia, hiponimia, meronimia, vinculación y causa.

3. Extendiendo las consultas de los usuarios

La Figura 1 muestra la arquitectura del sistema MESIA (la zona sombreada engloba el módulo de Generación de la Consulta Formal descrito en este artículo). Este componente incluye dos tareas principales.

3.1 Generación de la consulta

La primera consiste en transformar la consulta del usuario en lenguaje natural (LN) en una consulta formal que el buscador pueda ejecutar. Estas consultas pueden ser oraciones o frases sencillas, como las que se muestran a continuación, de las que hay que extraer los términos significativos para la búsqueda (actualmente sólo se manejan nombres y adjetivos):

- Becas para estancias en el extranjero
- Estancias en universidades extranjeras
- Proyectos de investigación financiados por la CAM para el año 2001
- Convocatorias de becas predoctorales
- Pruebas de acceso a la universidad

La extracción de términos significativos es un proceso de *shallow parsing* basado en la utilización de ARIES para etiquetar morfológicamente (categoría y rasgos de género y número) cada término de la consulta en LN realizada por el usuario. El predicado utilizado es: w(Lema,Categoria,Genero_Persona,Numero_Tense, Palabra,[]). Por ejemplo, para la obtención de los rasgos morfológicos del término *doctores*, se tiene:

?-w(L,C,G,N,doctores,[]).

L = doctorar, C = v (verbo)
G = sing_2
N = pres_subj;

L = doctor, C = n (nombre)
G = mas
N = sing.

Posteriormente, se lleva a cabo una segmentación de la consulta con el fin de detectar sintagmas nominales, preposicionales y verbales. El conocimiento lingüístico de la sintaxis de estos sintagmas lo forman un conjunto de autómatas en cascada representados como Redes de Transición Recursiva, que describen los distintos tipos de sintagmas básicos, Martínez (1998), Martínez y García-Serrano (1998), como el que se muestra en la Figura 2.

Esta segmentación de la consulta original en frases va a permitir obtener los núcleos y modificadores de los sintagmas para después aplicar una estrategia de colocación de los operadores booleanos (AND y OR) en la generación de la consulta formal.

3.2 Extensión de la consulta

La segunda funcionalidad consiste en extender los términos significativos de la consulta (formal) utilizando conocimiento lingüístico; para ello se añaden a los términos significativos de la consulta (enlazados con AND) las variantes morfológicas y semánticas mediante OR con el fin de construir una consulta en forma normal conjuntiva. Esta ampliación se hace de dos formas:

1. Utilizando la base de datos léxica ARIES se incluyen para cada término significativo sus variaciones morfológicas (por ejemplo, singulares y plurales) que pasarán a formar parte de la consulta formal. Una solución mejor consiste en obtener las raíces de los términos y que sólo formara parte de la consulta la raíz pero esta solución no es viable puesto que el Buscador de la CAM no realiza el mismo proceso para obtener los términos de indexación de los documentos.

   Por ejemplo, las variantes morfológicas del nombre *doctor* se obtienen con el predicado:

   ?-w(doctor, n, _ , _ , Palabra,[]).

   Palabra = doctoras; Palabra = doctores;
   Palabra = doctora; Palabra = doctor.

2. Utilizando la base de datos léxica EuroWordnet se amplia la consulta con términos sinónimos o semánticamente relacionados. Así, por ejemplo, si el usuario pregunta por *becas*, se incluye el sinónimo *ayudas*.

La integración de la base de datos léxica EuroWordnet se ha realizado mediante una conversión al lenguaje PROLOG de parte de la red semántica (sólo las relaciones de sinonimia, hiponimia e hiperonimia). El predicado utilizado para obtener los términos semánticamente

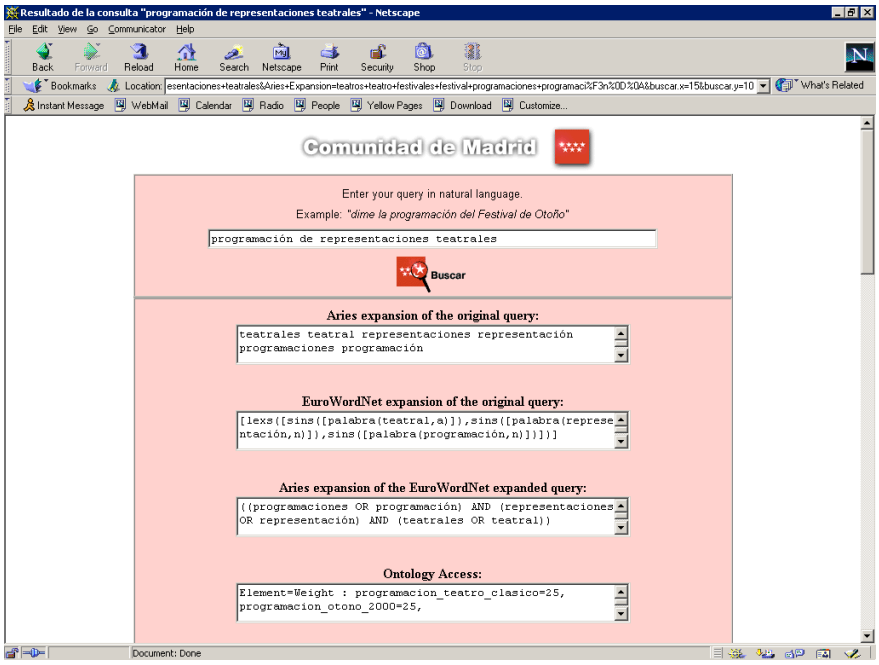

Figura 3: Interfaz del prototipo MESIA

relacionados a uno dado es:

ewn(Lexema, Categoria, sins(Sinonimos), hyper(Hiperonimos), hypo(Hiponimos))

/%Lexema: palabra de entrada a la red semántica
/% Categoría: nombre, adjetivo o verbo
/% Sinonimos:palabras con relación de sinonimia
/% Hiperonimos: conceptos más generales
/% Hiponimos: conceptos subclase

Para la expansión de los términos de la consulta, interesan únicamente las relaciones de sinonimia e hiponimia. Por ejemplo, para el término estancia, EuroWordnet devuelve los siguientes valores:

?- ewn(estancia, n, Sins, Hypers, Hypos).
Hypers= hyper([‘actividad humana’]),
Hypos = hypo([escala, estancia]),
Sins= sin([permanencia]) ? ;

Hypers= hyper([estancia]),
Hypos= hypo([visita]) ,
Sins = sin([ ]) ? ;

La Figura 3 muestra el interfaz de MESIA con una consulta de ejemplo.

4. Mejorando la presentación de resultados

En algunas ocasiones podemos suponer que si el usuario necesita información sobre un determinado tema, podría estar interesado en otros temas relacionados. Estos temas relacionados no están necesariamente descritos por las mismas palabras clave que el tema original; en este caso, la búsqueda basada en palabras clave no puede mejorar la recuperación aún incluso después de la expansión lingüística. Por otro lado, en la ordenación de las páginas que la consulta devuelve como resultado generalmente, los buscadores no consideran criterios adecuados aparte de los basados en datos estadísticos o históricos sobre consultas anteriores. Debido a ello, si se utiliza información del dominio de la colección de páginas sobre las que se trabaja, el tema de una consulta puede deducirse y los resultados pueden expandirse con temas similares sin considerar palabras clave sino conocimiento del dominio almacenado en una ontología.

En el proyecto MESIA se ha utilizado una ontología de un dominio específico consistente en una estructura jerárquica en forma de árbol cuyos nodos reflejan conceptos representados en las páginas Web del dominio. Cada nodo tiene asociado una descripción, un conjunto de palabras clave y una lista de enlaces relationados (*links*). Por ejemplo, el nodo “Arte sacro” tiene asociado las palabras “música”, “teatro”, “danza” y “religiosa”.

La utilización de la ontología podría haberse pensado como un paso más en la expansión de la consulta pero en MESIA se ha seguido el siguiente enfoque: tanto la búsqueda en Web como el acceso a la ontología se ejecutan simultáneamente y ambos utilizan la consulta ampliada como entrada. La ventaja de tener estos dos recursos trabajando independiente consiste en que si uno de ellos nos está disponible en el momento de la consulta, entonces el otro permanece operativo y el usuario siempre obtiene resultados. Cuando ambos procesos finalizan, se combinan los resultados con el fin de presentarlos al usuario. La clasificación de los resultados se realiza de acuerdo a la cercanía de cada link con la consulta del usuario; esta información se recupera de la ontología. La figura 4 muestra los links resultado de una consulta ejemplo (“programación de representaciones teatrales”) agrupados según las categorías proporcionadas por la ontología una vez se ha llevado a cabo la ordenación.

En una primera aproximación, el acceso a la ontología se gestiona mediante un sistema de pesos. Cada peso es un valor numérico que mide la relevancia de un nodo de la ontología para una consulta dada. Si una de las palabras clave asociada a un nodo aparece en la consulta del usuario, el nodo incrementa su peso y los nodos cercanos en la estructura (padres e hijos) consiguen también cierto peso. En cada acceso, este proceso se repite para cada palabra clave y cada nodo de la ontología. Una palabra clave puede estar asociada a distintos nodos por lo que todos esos nodos también incrementan sus pesos. Cuando el proceso de asignación de pesos concluye, todos los nodos que han adquirido pesos se ordenan de acuerdo a sus valores numéricos con el fin de visualizar los links que incorporan.

Aunque la ontología se construyó manualmente, la clasificación de las páginas de acuerdo a esta ontología podría hacerse de manera automática si se emplean el campo *metadata*; dado un campo metadata que contiene términos descriptivos del documento web, estos términos representarían una consulta y el documento se clasifica en el nodo de la ontología que obtiene el peso mayor.

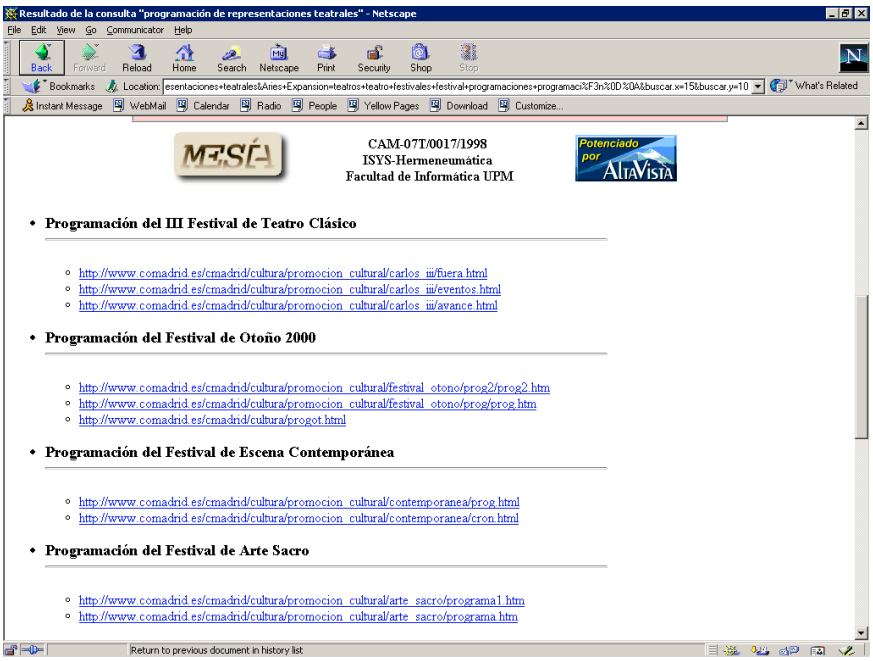

Figura 4: Clasificación de resultados de acuerdo a la ontología del dominio

5. Otros trabajos relacionados

En Staab et al. (1999) se describe el sistema de acceso a información turística en Web en inglés GETESS. Utiliza la semántica de los documentos (en un dominio restringido) que se encuentran en Web (bien porque se proporciona explícitamente bien porque se ha podido inferir utilizando algún sistema PLN incompleto); además utiliza métodos sintácticos de recuperación si los métodos en el ámbito semántico fallan. En particular, el sistema de búsqueda se caracteriza por: poseer conocimiento semántico que soporta la tarea de recuperación de información; comprender LN de forma parcial aunque robusta; permitir varias formas de interacción naturales al usuario y combinar conocimiento procedente de documentos estructurados y semiestructurados con sistemas de BD relacionales.

Fensel et al. (1999) y Chiang et al. (2000) utilizan una ontología (especificación consensuada y formal de un vocabulario utilizado para describir un dominio específico) para búsquedas en Web, pero a diferencia de MESIA basa la recuperación de información en una extensión del lenguaje HTML para incluir conocimiento semántico en la estructura de los documentos. En el trabajo de Möller et al. (1998), se utiliza también una ontología en un determinado dominio utilizando Description Logics (DL) para realizar recuperación de información así como en Todirascu et al. (2000).

En Julian et al. (1999) se propone una arquitectura multiagente de dos niveles para la gestión de la información existente en la red. En el primer nivel de la arquitectura se encuentran los agentes personales que asisten a los usuarios concretos mientras que en el segundo nivel, conectados por la red con los anteriores, están los agentes especializados en un tema concreto; para llevar a cabo la tarea en la que son expertos se conectan a las fuentes de información que se encuentran en la red (búsqueda de información, gestión de una BD, etc.).

6. Algunos trabajos de experimentación

Con el fin de evaluar cómo el conocimiento lingüístico afecta a los resultados de la recuperación de información, se ha llevado a cabo un primer estudio, García-Serrano y Martínez (2001), con siete consultas de usuario cortas (2-12 palabras/consulta) empleando cuatro tipos de experimentos para medir los valores de *precisión*, Baeza-Yates y Ribeiro-Neto (1999), teniendo en cuenta los 20 documentos mejor clasificados. Las consultas se han ejecutado utilizando la búsqueda avanzada proporcionada por Altavista. Las principales características de estos experimentos son:

- Experimento 1 (básico): consulta booleana con los términos relevantes combinados con el operador AND.
- Experimento 2 (Expansión con ARIES): cada término relevante expandido con sus variaciones de género, número, etc. combinados con operadores OR.
- Experimento 3 (Expansión con EWN): los términos relevantes se enriquecen con sinónimos/hiperónimos conectados por operadores OR.

- Experimento 4 (Expansión ARIES and EWN): incluye las expansiones de los experimentos 2 y 3.

Como hipótesis básica del trabajo se consideró que aunque la expansión con EWN añadía también los sinónimos no pertenecientes al contexto de la consulta, la propia conjunción de los términos iba a eliminar aquellos documentos *espúreos* (es decir, el propio proceso de recuperación realiza la tarea de desambigüación). El resultado de este primer estudio concluye que el uso de la morfología de ARIES generalmente mejora los resultados de la búsqueda. Sin embargo, la extensión con sinónimos/hiperónimos, aunque contribuye a recuperar documentos que no se recuperaban en el experimento básico, afecta a los valores de la precisión. La Tabla 1 muestra los resultados del estudio.

Básico	ARIES	EWN	ARIES+EWN
0,558	0.683	0.377	0.454

Tabla 1: Media de los valores de precisión

En un segundo estudio se han ejecutado 20 consultas cortas (3-10 palabras/consulta) utilizando los cuatro experimento anteriores con las siguientes modificaciones: cuando un termino relevante de la consulta es un verbo sólo se añade como variante morfológica la horma verbal en infinitivo correspondiente; en cuanto a las variatnes léxicas, sólo se han considerado los sinónimos extraídos de EWN teniendo en cuenta además la categoría gramatical (nombre, verbo, adjetivo, etc.) del término relevante; finalmente, se ha utilizado el campo *Order by* proporcionado por Altavista con los términos relevantes originales de la consulta con el fin de obtener los documentos relevantes mejor clasificados. Además, también se midió la precisión considerando los primeros 20 documentos ya que los usuarios sólo consideran los documentos iniciales resultado de la búsqueda. La Tabla 2 muestra los resultados de este segundo estudio que muestran que una combinación de rasgos morfológicos y semánticos mejora la recuperación de información. Sin embargo, en un examen de algunas consultas por separado se observa que la expansión con EWN afecta a la precisión.

Básico	ARIES	EWN	ARIES+EWN
0,551	0.667	0.685	0.779

Tabla 2: Media de los valores de precisión

7. Conclusiones y trabajos futuros

Este trabajo tiene como objetivo mostrar como la utilización de conocimiento lingüístico y del dominio mejora la recuperación de información y la interacción del usuario con la Web, actuando sobre las consultas del usuario. El metabuscador MESIA amplía la búsqueda habitual (consulta) con nuevas capacidades semánticas obtenidas a partir del análisis de la estructura de las páginas, del tratamiento lingüístico de algunas de las unidades de texto seleccionadas automáticamente y de la experiencia de uso.

En algunos casos, la expansión con EWN debe restringirse si se quieren mantener valores altos de precisión. Sin embargo, EWN plantea algunos problemas en el campo de la recuperación de información, Gonzalo et al. (1998a), Mihalcea (1999): por ejemplo, EWN no incluye relaciones semánticas con referencias cruzadas entre categorías gramaticales; las distinciones semánticas son demasiado sutiles en algunos casos (el grado de granularidad que se necesita es dependiente de la tarea); ausencia de información dependiente del dominio (sería muy interesante gestionar diversos dominios que permitieran almacenar preferencias semánticas dependiendo del dominio seleccionado).

Además, pueden establecerse contribuciones importantes, desde el punto de vista de los sistemas interactivos, pensadas para usuarios esporádicos que tienden a formular consultas cortas; una propuesta adecuada es ofrecer al usuario la posibilidad de especificar los términos relacionados con la consulta a partir de los derivados de EWN y, de esta manera, preparar diferentes consultas ampliadas que se ofrecerán al usuario.

Como mejoras futuras proponemos utilizar información sobre pesos en los términos de la consulta (dependiendo de si el sistema de recuperación las acepta) así como nuevos operadores léxicos tales como párrafo, oración, etc. tal y como se describe en Mihalcea (1999).

Otros problemas que debe solventar la tecnología de los motores de búsqueda en Internet relacionados con el resultado de la búsqueda son:

1. El problema de la *realimentación por relevancia* consistente en marcar documentos relevantes y no relevantes, o marcar cada documento con una jerarquía de relevancia (muy, poco, nada relevante) y ordenar los documentos por la relevancia que tienen según el usuario.
2. El usuario no emplea los términos de búsqueda adecuados y como los resultados de la búsqueda se centran más en la cantidad que en la calidad, las herramientas deberían manejar modelos de usuario para guiar el proceso o al menos ordenar los resultados.
3. En los sistemas de recuperación de información las variables que intervienen (índices, funciones de similitud, medidas de relevancia, etc.) se

refieren a medidas que no tienen que ver con el usuario. Se debería utilizar información sobre:

- estado del usuario
- necesidades del usuario (información relevante/irrelevante)
- filtro sobre las decisiones y percepciones del usuario (objetividad vs subjetividad).

Para la incorporación al sistema de un modelo de usuario en el que el sistema deberá clasificar el usuario pero no mediante una consulta exhaustiva a través de formularios, se utilizará un modelo parcial inicial, que se vaya modificando a través de la experiencia de uso del sistema por los diferentes usuarios junto con un sistema de mantenimiento gestión automática de los modelos de usuarios personales y genéricos.

Por otro lado, los trabajos actuales con el estándar XML permitirán que la tecnología de buscadores mejore considerablemente pues se podrán realizar búsquedas por contenido semántico en las páginas Web.

Agradecimientos

Queremos agradecer a Pablo Sánchez y Alberto Ruiz Cristina su colaboración en el proyecto de investigación PB48-0674-C04-04.

Referencias

Baeza-Yates, Ribeiro-Neto (1999), Ricardo Baeza-Yates, Berthier Ribeiro-Neto: Modern Information Retrieval. Chapter 3. Addison Wesley, 1999.

Bueno et al. (1999), F. Bueno, D. Cabeza, M. Carro, M. Hermenegildo, P. López, and G. Puebla (1999) ``The Ciao Prolog System: A Next Generation Logic Programming Environment, REFERENCE MANUAL" The Ciao System Documentation Series Technical Report CLIP 3/97.1, The CLIP Group School of Computer Science Technical University of Madrid.

Chiang et al. (2000), Chiang, R., Chua, C. and Storey V. A Smart Web Query Engine for Semantic Retrieval of Web Data. NLDB 200, Versalles, Francia, Junio 2000.

Fensel el al. (1999), Fensel, D., Angele, J., Decker, S., Erdmann, M, Schnurr, H., Staab, S., Studer, R. and Witt, A. On2broker: Semantic-based access to information Sources at the WWW. Proceedings of the World Conference on the WWW and Internet (WebNet 99). Honolulu, USA. Octubre 1999.

García-Serrano y Martínez (2001), García-Serrano, A., Martínez, P. An interface agent with linguistic skills. Proc. NLDB'01.

Gonzalo et al. (1998a), J. Gonzalo, F. Verdejo, I. Chugur and J. Cigarran, Indexing with WordNet synsets can improve Text Retrieval, Proc. COLING/ACL '98 Workshop on Usage of WordNet for NLP, Montreal, 1998.

Gonzalo et al. (1998b), J. Gonzalo, M.F. Verdejo, I. Chugur, F. López, A. Peñas. "Extracción de relaciones semánticas entre nombres y verbos en EuroWordNet". Actas de la SEPLN98, 1998.

Goñi et al. (1997), Goñi, J. M., González, J. C. y Moreno, A. ARIES: A lexical platform for engineering Spanish processing tools. Natural Language Engineering, 3 (4), pp. 317-345.

Julian et al. (1999), Julian, V., Carrascosa, C. y Soler, J. Una arquitectura de sistema multi-agente para la recuperación y presentación de la información. IV Congreso ISKO-España EOCONSID'99, 22-24 de Abril, Granada, 1999.

Liddy (1998), Liddy, E. D. Enhanced Text Retrieval Using Natural Language Processing. ASIS Bulletin, Abril/Mayo, V. 24, N. 4, 1998 (www.asis.org/Bulletin/Apr-98).

Martínez (1998), Martínez Fernández, P. Propuesta de estructuración del conocimiento lingüístico para interpretación de textos: Aplicación al diseño de bases de datos. Tesis Doctoral. Facultad de Informática, UPM, Julio 1998.

Martínez y García-Serrano (1998), Martínez, P. and García-Serrano, A. A Knowledge-based Methodology applied to Linguistic Engineering. In R. Nigel Horspool Ed., Systems Implementation 2000: Languages, Methods and Tools. London: Chapman & Hall, pp. 166-179, 1998.

Mihalcea (1999), R. Mihalcea, Word Sense Disambiguation and its application to Internet search, Master Thesis School of Engineering and applied Science, Southern Methodist University, 1999.

Miller (1995), Miller, G. A. WordNet: A lexical Database for English. Communications of the ACM, 38, 11, pp. 39-41, Noviembre 1995.

Möller et al. (1998), Möller, R., Haarslev, V. and Neumann, B. Semantics-based information retrieval. Information Technology and Knowledge Systems (IT & KNOWS). Viena y Budapest, 1998.

Palomar et al. (2001). PHORA: A NLP system for Spanish. CICLing 2001.

Staab et al. (1999), Staab, S., Braun, C., Bruder, I., Düsterhöft, A., Heuer A. A System for Facilitating and Enhancing Web Search. Proceedings of International Working Conference on Artificial and Neural Networks (IWANN '99). Alicante, ES, 1999,.

Todirascu et al. (2000), Todirascu, A., Beuvron, F, Galea, D. , Keith, B. and Rousselot, F. Using Semantics for Efficient Information Retrieval. NLDB 2000, Versalles, Francia, Junio 2000.

Voorhees (1999), Voorhees, E. Natural Language Processing and Information Retrieval. En Information Extraction: Towards Scalable, Adaptable Systems. Maria Teresa Pazienza (Ed.). LNAI Tutorial, Springer Verlag, 1999.

Vossen (1997), Vossen, P. EuroWordNet: a multiingual database for information retrieval. DELOS workshop on Cross-language Information Retrieval, Zurich, 1997.