Universidad Nacional de Rosario

Facultad de Ciencias Exactas, Ingeniería y Agrimensura

Escuela de Ingeniería Electrónica

Departamento de Sistemas e Informática

Tesis de Doctorado en Ingeniería

Mejora de la Recuperación de Información en Bases de Datos de Texto utilizando Recursos Lingüísticos

Claudia Deco

Director de tesis: Dra. Zulema Solana

Abril 2009

Resumen

Al convertirse la web en el mayor repositorio de conocimiento y en un medio de publicación fácilmente accesible para todos, la Recuperación de Información ha dejado de ser un campo exclusivo de los especialistas en Ciencias de la Información y ha pasado a ser un campo relacionado con cualquier persona. El maximizar la cantidad de documentos relevantes obtenidos para una consulta depende de la destreza de este especialista para preparar una estrategia de búsqueda adecuada. Si bien los usuarios no tienen por qué conocer técnicas de recuperación de información, la propuesta de esta tesis es mejorar los resultados de su búsqueda por medio de un "especialista" que implementa estas técnicas.

Se propone el refinamiento semántico de los conceptos de la consulta a fin de mejorar la precisión de los resultados, utilizando recursos lingüísticos para construir una estrategia de búsqueda adecuada. El refinamiento semántico propuesto consiste en: guiar al usuario para desambiguar los conceptos ingresados por él, permitirle seleccionar conceptos jerárquicamente relacionados a fin de precisar los documentos a recuperar, y expandir semánticamente los conceptos a fin de aumentar la cantidad de documentos a recuperar. Los recursos lingüísticos que pueden utilizarse son tesauros, diccionarios, diccionarios multilinguales y ontologías. Qué recursos utilizar, depende del área del conocimiento de la consulta y de los recursos disponibles para ese área.

Se evalúa el refinamiento semántico, eligiendo el recurso WordNet para consultas de dominio general y el recurso MeSH, especializado en el área salud, para las consultas en un dominio específico. Las experiencias realizadas muestran que aumenta la precisión de los resultados en un 19,03 % en el dominio general y en un 33,50 % en un dominio específico del conocimiento.

Otro aspecto tratado es la inmersión del refinamiento semántico en un motor de búsqueda propio de un sitio web y en sistemas de recomendación. Los resultados experimentales muestran que el uso del refinamiento mejora las prestaciones del motor de búsqueda del sitio, con respecto a su uso en su forma estándar,

obteniéndose un incremento cercano al 33% en la precisión y duplicando aproximadamente la cantidad de documentos recuperados. Para la inclusión del refinamiento en sistemas de recomendación se elige el área educación, planteando el agregado de la personalización de los resultados utilizando metadatos del usuario y metadatos de los documentos. De esta forma se potencia la recuperación obtenida del refinamiento semántico porque se ordenan los resultados de distinta forma según el usuario y el momento en que éste haya realizado la consulta.

Abstract

As the Web has become one of the biggest repository of knowledge easily accessible for everyone, the Information Retrieval has stopped to be an exclusive field of information sciences specialists, and it has become a field related with any person. Although users do not have to know information retrieval techniques, the proposal of this thesis is to improve query results by using a "specialist" that implements these techniques. For this, a semantic refinement is proposed. This semantic refinement acts as a specialist in information sciences and prepares, using linguistic resources, an appropriate search strategy that represents the user's information need. The semantic refinement consists on three steps. First, it guides the user in the disambiguation of the words submitted by him. Then, it allows the user to select concepts hierarchically related in order to reduce the amount of documents to retrieve. Finally, it expands semantically concepts to increase the amount of documents to be retrieved. Linguistic resources that can be used are thesauri, dictionaries, multilingual dictionaries and ontologies. What resources are used, depends on the domain of knowledge, and on resources available for that domain.

Semantic refinement is evaluated by using WordNet as a linguistic resource for information retrieval in a general domain knowledge. Also, it was evaluated in a specific domain, by using MeSH, as a linguistic resource for health. Experimental results show that precision increases in a 19.03% in a general domain, and in a 33.50% in a specific domain of knowledge. Semantic Refinement was also applied in a website's search engine. Experimental results show that the refinement improves the performance of the site's search engine in almost 33% in precision and approximately doubles the number of documents retrieved. Another issue addressed in this thesis, is the immersion of the proposed refinement, in an recommender system of educational materials for a personalized retrieval. Personalization consists in ordering the results according to the user's profile depending on his/her preferences. The use of the refinement inside the recommender system improves the semantic search, and because of this the recommendation will be improved.

Agradecimientos

Quisiera agradecer a Dios por estar conmigo en cada paso que doy y por ser mi soporte en todo momento y a mi esposo José, mis hijos María Belén y Juan Pablo, y mis abuelos Octavia y Luis, porque gracias a todos ellos soy lo que soy y con cuyo apoyo he logrado llegar hasta aquí.

También quisiera agradecer a todas y cada una de las personas que me han brindado todo el apoyo, colaboración y ánimo, especialmente a los miembros del Departamento de Sistemas e Informática de esta Facultad, del Grupo Infosur de Lingüística de la Facultad de Humanidades, y del Departamento de Investigación Institucional de la Facultad de Química e Ingeniería de la Universidad Católica Argentina.

En particular, quisiera agradecer a mi directora, la Profesora Dra. Zulema Solana, de la Facultad de Humanidades, quien me guió durante todo el proceso de investigación y escritura de esta tesis, y de quien recibí valiosos aportes en todas las instancias de este trabajo. Hago extensivo este agradecimiento a la Dra. Regina Motz y a mi compañera del grupo de investigación de la universidad, la M. Sc. Cristina Bender.

Índice

Lista de Figuras

Lista de Tablas

Capítulo 1: Introducción

1.1 El problema

La forma tradicional de búsqueda de información obligaba a un usuario a recorrer biblioteca por biblioteca y consultar cada uno de sus ficheros para satisfacer su necesidad de información. Este problema se solucionó con la aparición, en la década del ’80, de las grandes bases de datos bibliográficas que reúnen toda la bibliografía especializada sobre un área del conocimiento. Estas bases de datos se consultaban en línea, en bibliotecas o centros de información, y con el apoyo de un experto en ciencias de la información, que era el encargado de transformar la necesidad de información del usuario en una estrategia de búsqueda adecuada. Una estrategia de búsqueda es una expresión lógica compuesta por distintos conceptos combinados con los conectores lógicos de conjunción, disyunción y negación.

En la década de los ’90, con la aparición de Internet y el abaratamiento de los costos de equipamiento, el usuario dejó de concurrir a las bibliotecas o centros de información y comenzó a buscar información por sus propios medios. Por lo tanto dejó de utilizar el apoyo del experto en ciencias de la información para expresar su necesidad de información. Como consecuencia, y agregando a esto la explosión de información disponible en la web, resulta muy difícil para el usuario encontrar eficientemente información útil, dado que no es capaz de preparar una estrategia de búsqueda adecuada. Además, el tiempo que éste perdía años atrás recorriendo bibliotecas, lo pierde ahora buscando en una y otra base de datos, y recorriendo páginas obtenidas a través de buscadores, en búsqueda de información útil.

Así, el exponencial crecimiento de información disponible en la web lleva al problema que los usuarios no son capaces de encontrar la información que buscan en una forma eficiente y simple, y frecuentemente no ven satisfechas sus necesidades de información.

La recuperación de información utilizando la web se puede realizar consultando bases de datos o empleando buscadores. Entonces, el usuario ingresa a una base de datos bibliográfica, y allí realiza su búsqueda. De esta manera obtiene

una primera lista de referencias bibliográficas sobre su tema de interés. Para ampliar estos resultados accede a otra base de datos, y vuelve a realizar su búsqueda en ésta, obteniendo otro conjunto de referencias. Así continúa con todas las bases de datos que conozca. Para obtener más información, recurre a consultar algunas páginas conocidas según su experiencia, tales como sitios de consensos, asociaciones internacionales y nacionales vinculadas al tema, etc. En cada caso encuentra algunas páginas que responden a su interés de búsqueda y otras páginas que no le son útiles. Como alternativa final, puede realizar su consulta a través de uno o más buscadores o metabuscadores de páginas web, obteniendo un conjunto de páginas, algunas de las cuales pueden estar relacionadas con su interés de búsqueda, y que luego debe recorrer una a una para ver si le son de utilidad o no.

Estas búsquedas, realizadas en diferentes bases de datos, tienen el problema de que cada una utiliza diferentes interfaces y diferentes términos de búsqueda. Además, las bases de datos utilizan distintas sintaxis para la escritura de la estrategia de búsqueda. Por esto, el usuario debería conocer las sintaxis correspondientes a cada una de las fuentes de información que consultará. Otro problema es que en el resultado de la consulta el usuario se encontrará con un gran número de documentos duplicados, debido a la redundancia de información existente en la web.

Por otro lado, el uso de buscadores para localizar los documentos de interés tiene el problema de necesitar descartar manualmente los documentos no relevantes para la búsqueda. Esto está asociado al problema de saber establecer con precisión la frase por la cual se buscará el documento. Por ejemplo, no tener en cuenta el uso de sinónimos al plantear la búsqueda puede reducir notoriamente el número de documentos a ser retornados por el buscador, o una frase de búsqueda incompleta puede retornar cientos de documentos totalmente fuera del dominio de la aplicación buscada. También son significativos el tiempo y el esfuerzo que le demandará a un usuario realizar la búsqueda explorando uno y otro lugar y revisando los resultados obtenidos en cada fuente.

Podemos entonces precisar que los problemas a los que se ven enfrentados los usuarios son básicamente dos: cómo especificar la consulta y cómo interpretar las respuestas obtenidas.

Además de estas cuestiones con los usuarios, otros desafíos para la búsqueda en la web se relacionan con las dificultades que se pueden presentar con los datos. Los problemas relacionados con los datos se refieren a su ubicación en forma distribuida, su calidad, la redundancia, la falta de estructura, la volatilidad y la heterogeneidad semántica y/o estructural [Baeza, 1998]. La naturaleza intrínseca de la web hace que los datos estén distribuidos en diferentes computadoras y plataformas. Respecto a la calidad de los datos, la web se puede considerar como un nuevo medio de publicación, pero en la mayoría de los casos no hay un proceso ni control editorial. Además los datos no tienen una estructura uniforme y casi el 30% de los documentos está duplicado [Shivakumar et al., 1998]. La volatilidad de los datos se debe a la dinámica de Internet, ya que las páginas pueden cambiar, aparecer o desaparecer en forma muy rápida. La heterogeneidad se presenta al tratar con múltiples tipos de medios: imágenes, videos, texto; y diferentes idiomas y alfabetos.

Si bien los usuarios no tienen por qué conocer técnicas de recuperación de información, se mejorarían los resultados de su búsqueda si se le sugiriera expandir o restringir los conceptos semántica y multilingualmente y así lograr que en su respuesta los documentos recuperados sean los documentos relevantes. La propuesta de esta tesis es mejorar la recuperación de información en la web mediante el uso de recursos lingüísticos adecuados, logrando la mejora de la precisión de los resultados. Además, se muestran resultados experimentales que verifican esta mejora.

1.2. Ejemplo motivador

Supongamos que un médico desea obtener bibliografía científica sobre “cáncer de pulmón”. Una posibilidad para obtener esta información es utilizar la web y a través de ella ingresar a bases de datos bibliográficas, páginas específicas de medicina, o buscadores.

Este usuario ingresará a una base de datos bibliográfica especializada en medicina, como por ejemplo Medline[1], y allí realizará su búsqueda, obteniendo una primera lista de referencias bibliográficas. Para ampliar esta búsqueda accederá a

otra base de datos, como por ejemplo Excerpta Medica[2], y volverá a realizar su búsqueda en ésta obteniendo otro conjunto de referencias bibliográficas. Así continuará con todas las bases de datos que conozca. Para obtener más información realizará su consulta a través de uno o más buscadores o metabuscadores de páginas web, obteniendo un conjunto de páginas, algunas de las cuales podrán estar relacionadas con su interés de búsqueda y que luego deberá recorrer una a una para ver si le son de utilidad o no.

Todo este proceso representa varios problemas para este usuario. Deberá conocer las distintas sintaxis necesarias para cada una de las fuentes de información que consultará: la sintaxis del lenguaje de consulta de la base de datos Medline es distinta a la de Excerpta Medica y ambas son distintas a la sintaxis de un buscador. Tendrá que descartar los resultados no relevantes a su búsqueda. Se encontrará con documentos duplicados: un documento puede estar en más de una base de datos y además ser recuperado por un buscador. En todos estos resultados únicamente encontrará documentos que contengan sólo la frase "cáncer de pulmón" si no tuvo en cuenta la existencia de sinónimos de este concepto en el planteo de sus búsquedas. Además, serán significativos el tiempo y el esfuerzo que le demandará realizar la búsqueda explorando uno y otro lugar y revisando los resultados obtenidos en cada fuente.

1.3. Abordaje propuesto

El objetivo general de esta tesis es mejorar la recuperación de información en la web mediante la utilización de recursos lingüísticos. Para esto se analiza la analogía entre la búsqueda en la web y la búsqueda en los sistemas de recuperación de información clásicos. Se analizan además los distintos recursos lingüísticos que pueden utilizarse para la preparación de una estrategia de búsqueda y se propone su utilización para el refinamiento semántico de los conceptos. Otro de los objetivos es

[1] Biblioteca Nacional de Medicina de Estados Unidos; www.nlm.nih.gov/medlineplus/
[2] Editorial Elsevier; www.excerptamedica.com

evaluar el desempeño de distintos recursos lingüísticos que se pueden usar para este refinamiento semántico.

Una forma de mejorar la estrategia de búsqueda es mediante una expansión de la consulta. No siempre un término representa en forma adecuada un concepto de interés para el usuario. Expandir la consulta es encontrar otros términos equivalentes o más adecuados para expresar un concepto correctamente. La consulta inicial, tal como es provista por el usuario, puede ser una representación inadecuada o incompleta de la información que éste necesita, ya sea en sí misma o en relación con la representación de las ideas en los documentos. Para seleccionar términos para la expansión se debe disponer de estructuras de conocimiento que sean independientes del proceso de búsqueda, tales como tesauros o diccionarios, tanto específicos del dominio como globales. Un ejemplo de una estructura de conocimiento específica del área salud es el Medical Subject Headings [MeSH], y una global es WordNet [WordNet].

Un problema en la expansión de consulta, es cómo definir cuáles términos están estrechamente asociados con los términos de la consulta. Esta expansión puede ser desarrollada manual, automática o semiautomática. La expansión de la consulta propuesta en este trabajo es semiautomática. La expansión semiautomática es interactiva dado que se le proponen al usuario términos de búsqueda como parte del proceso de reformulación de la consulta. En este tipo de expansión hay una responsabilidad conjunta entre el sistema y el usuario en la selección de términos para la expansión. Por un lado, el sistema sugiere términos y los presenta al usuario; y por el otro, son los usuarios quienes seleccionan términos de acuerdo a sus preferencias. Entonces, es el usuario quien toma la decisión final sobre la utilidad de un término. Por lo tanto, son las elecciones de términos del usuario, las que reflejan la importancia relativa y la utilidad de términos desde la perspectiva del usuario, incrementando de esta forma su satisfacción. Así, la arquitectura propuesta sugiere términos y es el usuario quien decide cuál o cuáles términos representan su interés de búsqueda.

Esta tesis se enmarca dentro de dos proyectos de investigación, actualmente en curso. El primero es el proyecto “INFOSUR: Investigación y Desarrollo” de la

Facultad de Humanidades y Artes (UNR) dirigido por la Dra. Zulema Solana [Solana, 2005-2008]. Y el segundo el proyecto “Recuperación de Información en Bases de Datos de Texto” de la Facultad de Ciencias Exactas, Ingeniería y Agrimensura (UNR) dirigido por la autora de esta tesis, acreditado para el período 2007-2010 [Deco, 2007-2010] y presentado en el Workshop de Investigadores en Ciencias de la Computación [Deco et al., 2008b]. Este segundo proyecto parte de resultados obtenidos en proyectos previos. Uno de estos proyectos es “Tecnologías Middleware e Internet: búsqueda asistida de evidencia clínica en medicina” [Plüss, 2004-2006], codirigido por la autora de esta tesis, durante el período 2004-2006. El otro proyecto previo es “Recuperación de información basada en semántica (RIBS)” [Deco, 2004-2006], dirigido por la autora de esta tesis, durante el mismo período.

La Figura 1.1 ilustra el abordaje general del proyecto RIBS.

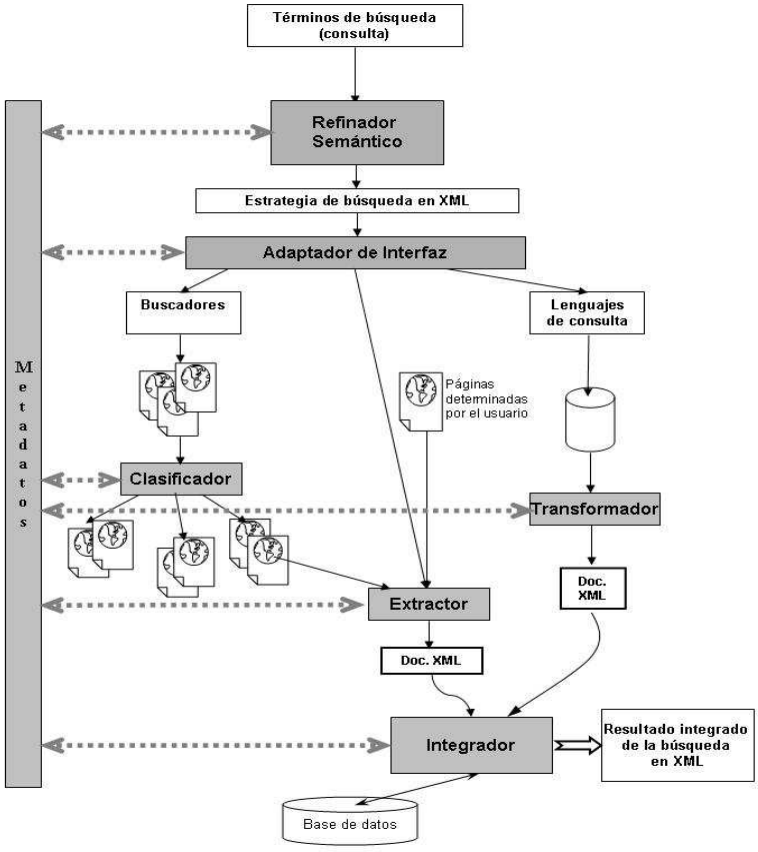


Figura 1.1: Abordaje general del proyecto RIBS.

En este contexto de trabajo, esta tesis focaliza uno de los módulos de esta arquitectura: el *refinador semántico*, el cual analiza los términos ingresados y construye una estrategia de búsqueda que represente la necesidad de información del usuario.

En este proyecto propuse una arquitectura para mejorar la recuperación de información en la web. En esta arquitectura, el usuario ingresa uno o varios términos de búsqueda, que son transformados por el módulo Refinador Semántico en una estrategia de búsqueda. Un término de búsqueda es una palabra o una frase que representa el interés de búsqueda de información del usuario. Una estrategia de búsqueda es una expresión lógica compuesta por distintos términos combinados con los conectores lógicos de conjunción, disyunción y negación (AND, OR y NOT respectivamente). Esta estrategia luego es convertida por el Adaptador de Interfaz a la sintaxis de cada una de las distintas fuentes. El módulo Clasificador agrupa las páginas resultantes de buscadores según criterios de clasificación. El resultado de la consulta enviada a cada una de las fuentes es convertido a XML (eXtensible Markup Language) por los módulos Extractor y Transformador, según corresponda. Finalmente, los resultados de cada fuente se integran en el módulo Integrador, para devolverle al usuario una única respuesta, la cual puede ser almacenada también en una base de datos. Como formato de intercambio entre los distintos módulos se adoptó XML por ser el formato estándar del Consorcio WWW [W3C]. Los Metadatos contienen información sobre los recursos lingüísticos disponibles, los DTDs (Document Type Definition) de los XML utilizados, criterios a utilizar para la clasificación y la integración, área del conocimiento de la consulta, etc.

El *Refinador Semántico* utiliza los recursos lingüísticos para la preparación de la estrategia de búsqueda. Qué recurso o recursos lingüísticos pueden utilizarse dependen del área del conocimiento y esta información también está en los metadatos.

La estrategia genérica de búsqueda producida por el refinador semántico es luego procesada por el *adaptador de interfaz* para adaptarla a la sintaxis de búsqueda de cada uno de los distintos tipos de fuentes existentes en la web que se deseen consultar. Dentro del proyecto RIBS los distintos tipos de fuentes que se consideran son bases de datos de texto y páginas web. Las páginas web pueden corresponder a

los resultados de una consulta realizada a través de un buscador, o pueden ser un sitio determinado conocido por el usuario. La diferencia entre las bases de datos y las páginas web es que las primeras tienen datos estructurados y las páginas web tienen datos no estructurados. Además, a las bases de datos de texto se accede utilizando un lenguaje de consulta propio del motor en el que están implementadas.

En la consulta a bases de datos, si los resultados devueltos por el motor no están en XML se utiliza el módulo *transformador* para convertir estos resultados a XML.

En el caso de páginas web obtenidas a través de una consulta en un buscador, como el número de enlaces obtenido puede ser muy grande y parte de la información puede no ser pertinente al interés del usuario, se agrega un *clasificador*. Este es el encargado de agrupar los documentos resultantes según la información que contengan sobre el tema y ciertos criterios predeterminados [Motz et al., 2003a], [Bender et al., 2004], [Bender et al., 2005]. Esto evita al usuario interesado en obtener bibliografía, la revisión, por ejemplo, de páginas comerciales. La salida de este clasificador es un conjunto de páginas web por cada categoría de clasificación.

Las páginas web resultantes son procesadas por un módulo *extractor*, que se encarga de dar estructura en formato XML a la información contenida en ellas a partir de criterios predeterminados por el interés del usuario. Esta extracción se realiza utilizando wrappers [Gruser et al., 1998], [Bender et al., 2004].

Finalmente, todos los documentos XML producidos por el transformador y el extractor, son unificados por el *integrador* para presentar una única respuesta al usuario. El integrador se encarga de homogeneizar las distintas estructuras provenientes de los distintos documentos XML en una única estructura. En esta integración se eliminan además los documentos duplicados y se efectúa un ordenamiento según un ranking de importancia de los documentos hallados a fin de presentarle al usuario los más relevantes primero.

Esta tesis se focaliza en el *refinador semántico*, el cual analiza los términos ingresados y construye una estrategia de búsqueda que represente la necesidad de información del usuario, utilizando para esto recursos lingüísticos.

El *refinamiento semántico* que se propone consiste en: guiar al usuario para *desambiguar* los conceptos ingresados por él, permitirle *seleccionar* conceptos

jerárquicamente relacionados a fin de precisar los documentos a recuperar y *expandir* semánticamente los conceptos a fin de aumentar la cantidad de documentos a recuperar.

Un problema que se presenta con respecto a la semántica es la *desambiguación* de conceptos. Basta un ejemplo muy simple como el hecho de buscar la palabra "cáncer" para comprobarlo. Cáncer puede referirse a la enfermedad, a la constelación de estrellas o al signo zodiacal. La desambiguación que realiza el usuario permite continuar el proceso, en las etapas siguientes, con sólo aquellos términos que están vinculados conceptualmente con la acepción de su interés de búsqueda. La solución propuesta es utilizar recursos lingüísticos, tales como diccionarios, tesauros u ontologías, para decidir dentro de qué contexto se está buscando el concepto ingresado por el usuario.

El objetivo de la *selección* de conceptos relacionados es mostrarle al usuario una jerarquía de conceptos vinculados con el concepto ingresado por él, a fin de que éste se reubique, si es necesario, en la jerarquía conceptual. De esta forma, el usuario puede refocalizar su búsqueda y así aumentar la precisión de los resultados obtenidos. Los tesauros y las ontologías son recursos lingüísticos que proveen estas jerarquías conceptuales.

El objetivo de la *expansión semántica* es recuperar documentos que también sean relevantes aún cuando no respondan rigurosamente a los términos utilizados por el usuario. Es decir, la expansión semántica consiste en incorporar a la búsqueda términos que sean conceptualmente equivalentes: sinónimos y términos relacionados. Por ejemplo, ante la búsqueda del término *padre*, se puede expandir semánticamente agregando su sinónimo *papá* y su término relacionado *madre*. La expansión del concepto puede hacerse utilizando recursos lingüísticos específicos del área del conocimiento disponibles en línea. Esta expansión también puede hacerse desde el punto de vista multilingual, utilizando en este caso diccionarios multilinguales.

Entonces, el refinamiento semántico, basándose en recursos lingüísticos, prepara una estrategia de búsqueda a partir de los conceptos ingresados por el usuario. Este refinamiento se hace en forma semiautomática, pues en ciertas tareas se requiere la participación del usuario. La desambiguación de los conceptos la realiza el usuario, seleccionando la acepción del concepto que corresponde a su interés de

búsqueda. La selección de conceptos jerárquicamente relacionados la realiza el usuario a partir de la jerarquía propuesta por el refinador. La expansión semántica se realiza en forma automática. El resultado es una estrategia de búsqueda preparada en forma automática por el refinador. Esto se detalla en la Sección 4.1.

El esfuerzo inicial que se pretende por parte del usuario en la desambiguación y en la selección de conceptos jerárquicos relacionados sugeridos por el sistema, será recompensado evitándole a posteriori la lectura y el descarte de los documentos que no sean de su interés, y mejorando así los indicadores de la Recuperación de Información.

1.4. Guía de lectura de la tesis

El resto de la tesis está organizada de la siguiente forma: en el Capítulo 2 se presentan conceptos básicos de recuperación de información y los recursos lingüísticos que se utilizan para preparar la estrategia de búsqueda para una consulta. En el Capítulo 3 se presentan trabajos relacionados. En el Capítulo 4 se propone una arquitectura para el Refinamiento Semántico, que permite armar la estrategia de búsqueda y se presenta un prototipo. En el Capítulo 5 se describen las experiencias realizadas en un dominio general del conocimiento y las realizadas en un dominio específico. En el Capítulo 6 se presenta la aplicación del Refinamiento Semántico a un motor de búsqueda de un sitio web. En el Capítulo 7 se presenta la inmersión del refinamiento dentro de un sistema de recomendación. Finalmente, en el Capítulo 8 se presentan las conclusiones y trabajos futuros. Se adjuntan a este documento tres apéndices, uno que incluye los relevamientos realizados de recursos lingüísticos disponibles en línea, en el segundo se discuten los problemas que se presentan en la búsqueda multilingual, y en el tercero se detalla el prototipo.

Capítulo 2: Conceptos básicos

2.1. Recuperación de Información

El objetivo principal de la Recuperación de Información es satisfacer la necesidad de información planteada por un usuario en una consulta en lenguaje natural especificada a través de un conjunto de palabras claves, también llamadas descriptores. En general, este proceso hacia la recuperación de documentos relevantes a la consulta presentada, no es un proceso simple debido a la complejidad semántica del vocabulario.

Según Baeza y Ribeiro, la *Recuperación de Información* o *Information Retrieval* (IR) es la representación, almacenamiento, organización y acceso a ítems de información [Baeza et al., 1999]. La representación y organización de los ítems de información no son un problema simple de resolver, al igual que la caracterización de la necesidad de información del usuario tampoco lo es.

En el modelo tradicional utilizado en la IR [Silberschatz et al., 1998], la información se organiza en documentos y se supone que existe un gran número de éstos. El proceso de recuperación consiste en localizar documentos relevantes a partir de la información aportada por el usuario. Un ejemplo típico de un sistema de recuperación de información son los catálogos interactivos de las bibliotecas, donde una entrada del catálogo es ejemplo de un documento.

Un usuario puede desear recuperar un documento concreto o un conjunto de éstos. Para la recuperación, los usuarios suelen describir los documentos deseados mediante un conjunto de palabras claves. Por ejemplo, se puede utilizar la palabra clave “cáncer de pulmón” para buscar información sobre este tema.

Los documentos tienen un conjunto de palabras claves asociado. Los sistemas de recuperación de información recuperan aquellos documentos cuyos conjuntos de palabras claves contengan las proporcionadas por el usuario.

La meta principal de un sistema de IR es recuperar información que podría ser útil o importante para el usuario, y no sólo datos que satisfagan una consulta dada [Manning et al., 2007].

Un sistema de recuperación de *datos*, tal como una base de datos relacional, trata con datos que tienen una estructura y una semántica bien definidas. Un sistema de recuperación de datos permite recuperar todos los objetos que satisfacen las condiciones especificadas en una expresión regular o en una expresión del álgebra relacional. Por ejemplo, si se consulta por la palabra “cáncer” un sistema de recuperación de este tipo recuperará solamente aquellos objetos que contengan exactamente dicha palabra.

Entonces, un sistema de recuperación de *datos* sólo recupera los datos que coinciden exactamente con el patrón ingresado por el usuario, mientras que un sistema de recuperación de *información* recupera datos relevantes que hagan la mejor coincidencia parcial con el patrón dado. Esto se debe a que la recuperación de información generalmente trata con texto en lenguaje natural, el cual no está siempre bien estructurado y podría ser semánticamente ambiguo. Por ejemplo, si se realiza una consulta por el término “cáncer” en un sistema de recuperación de información, además de obtener como resultado los documentos que contengan este término, se debería obtener también los documentos en que aparezca ”neoplasma”, ”carcinoma”, “cancerígeno”, etc..

Una *consulta* en un sistema de recuperación de información es una solicitud de documentos pertenecientes a algún tema. Dada una colección de documentos y una consulta del usuario, el objetivo de una *estrategia de búsqueda* es obtener todos y sólo los documentos relevantes a la consulta. El problema central se reduce a establecer una correspondencia entre los términos utilizados en la consulta y los términos del documento.

En la estrategia de búsqueda se utilizan operadores lógicos (*y*, *o*, *no*) para vincular los términos ingresados. De esta forma, el usuario puede expresar distintas necesidades de información. Por ejemplo, puede solicitar los documentos que contengan las palabras “cáncer” y “quimioterapia”, si desea recuperar información sobre tratamientos del cáncer con quimioterapia. Si deseara recuperar información sobre alguno de estos temas, pero no necesariamente juntos en el mismo documento,

lo expresaría ingresando: “cáncer” o “quimioterapia”. Y si quiere descartar la quimioterapia como tratamiento del cáncer, solicitaría documentos que contengan la palabra “cáncer” pero no “quimioterapia”.

La Recuperación de Información es una tarea compleja porque se enfrenta con varios problemas. Por un lado, los autores y los usuarios frecuentemente utilizan diferentes palabras o expresiones cuando se refieren a un mismo concepto. Por ejemplo, en medicina, “cáncer” puede también ser expresado como “neoplasma”.

Si en un documento, en lugar del término “cáncer” apareciera la palabra “neoplasma”, este documento no se recuperaría. Este problema se puede resolver haciendo uso de sinónimos. Cada palabra puede tener definido un conjunto de sinónimos y la aparición de una palabra puede sustituirse por la disyunción de todos sus sinónimos, incluyendo la propia palabra. Por lo tanto, la consulta:

“quimioterapia” y “cáncer”

puede sustituirse por:

“quimioterapia” *y* (“cáncer” *o* “neoplasma”)

Por otro lado, algunos términos pueden tener significados diferentes. Por ejemplo, la palabra “cáncer” puede referirse a una enfermedad (en medicina), a un signo zodiacal (en astrología) o a una constelación de estrellas (en astronomía). En este caso, se debe desambiguar el término, para indicar a qué área del conocimiento pertenece. La desambiguación se puede hacer agregando otros términos específicos relacionados con la acepción de interés. Por ejemplo, utilizar (“cáncer” y “terapia”) en lugar de usar sólo el término “cáncer”, si se lo desea recuperar en el área medicina. Otra forma de realizar la desambiguación es utilizando tesauros, cuya descripción se encuentra en la Sección 2.4. Los tesauros indican un término alternativo a utilizar en reemplazo del inicial para desambiguar el término; por ejemplo, la utilización del término “neoplasma” en lugar del término original “cáncer”.

Otro problema importante de la recuperación de información es el grado de relevancia de los documentos. Un sistema de recuperación de información para ser efectivo, debe en alguna forma interpretar los contenidos de los documentos en una colección y ordenar los resultados según un ranking de acuerdo al grado de

relevancia que tenga respecto a la consulta del usuario. Por lo tanto, la noción de relevancia es el centro de la recuperación de información. Así, el objetivo de la IR es recuperar todos los documentos que sean relevantes a una consulta del usuario y recuperar la mínima cantidad de documentos no relevantes.

En la recuperación de información existe además la figura del *especialista en ciencias de la información* que es el encargado de expresar la necesidad de información del usuario en una estrategia de búsqueda. El maximizar la cantidad de documentos relevantes obtenidos para esta consulta depende de la correcta preparación de esta estrategia de búsqueda.

Al convertirse la web en el mayor repositorio de conocimiento y en un medio de publicación fácilmente accesible para todos, resurgió la recuperación de información que hasta ahora era sólo usada por bibliotecarios y especialistas en ciencias de la información. Por lo tanto, en la última década, la recuperación de información ha dejado de ser un campo exclusivo de estos especialistas de la información y ha pasado a ser un campo relacionado con cualquier persona.

2.2. Recuperación de Información en la Web

Los motores de búsqueda recolectan páginas de la web, las indexan, buscan en los índices las palabras claves ingresadas en la consulta, utilizan algoritmos de ranking para ordenar los resultados y muestran al usuario los documentos resultantes.

Una página web corresponde a un documento en la recuperación de información tradicional. La recuperación de información en la web considera como una colección de documentos la parte de la web que está públicamente indexada, excluyendo las páginas que no puedan ser indexadas por ser muy dinámicas o por ser privadas.

Los desafíos para la búsqueda en la web se relacionan con problemas que se pueden presentar con los datos y con problemas que se pueden presentar con los usuarios [Baeza et al., 1999] [Allan et al., 2002].

Los principales desafíos en la recuperación de información en la web respecto a los datos, son:

- Datos distribuidos: debido a la naturaleza intrínseca de la web, los datos están expandidos en diferentes computadoras y plataformas.
- Datos volátiles: debido a la dinámica de Internet, los datos tienen un alto porcentaje de volatilidad. Los documentos pueden cambiar o desaparecer en forma muy rápida, o se pueden agregar nuevos documentos y/o computadoras.
- Gran volumen: el crecimiento exponencial de la web lleva a tener millones de documentos.
- Datos no estructurados y redundantes: no hay una estructura uniforme en los datos y casi el 30% de los documentos está duplicado [Shivakumar et al., 1998].
- Calidad de los datos: la web es un nuevo medio de publicación, pero en la mayoría de los casos no hay un proceso ni control de editorial. Por lo tanto los datos pueden ser falsos, ser no válidos por no ser actuales o estar pobremente escritos y con errores.
- Datos heterogéneos: la heterogeneidad se presenta al tratar con múltiples tipos de medios: imágenes, videos, texto; y diferentes idiomas y alfabetos.

El problema con la distribución de los datos se puede resolver enviando la consulta a los distintos repositorios de información e integrando los resultados. La calidad de la información proveniente de páginas web puede resolverse clasificando el origen de las fuentes. La redundancia, eliminando los duplicados en la integración. El problema de la falta de estructura se resuelve estructurando los documentos a un estándar de intercambio de datos. La heterogeneidad estructural integrando las estructuras mediante un estándar de intercambio de datos, y la heterogeneidad semántica integrando la información mediante recursos lingüísticos adecuados.

Los problemas encarados por los usuarios son cómo especificar la consulta y cómo interpretar las respuestas obtenidas.

Estos problemas se pueden solucionar con el refinamiento semántico propuesto en esta tesis. Este refinamiento actúa como un “especialista” en ciencias de la información, con el objetivo de preparar, con la utilización de recursos lingüísticos

apropiados, una estrategia de búsqueda adecuada a la necesidad de información del usuario.

Algunos motores de búsqueda están potenciados por técnicas de recuperación de información, pero cubren sólo un 25% a un 55% de la web y la mayoría está en inglés [Tsikrika, 2001]. Algunos ejemplos de estos motores de búsqueda son: AltaVista (www.altavista.com), Excite (www.excite.com), Google (www.google.com), Yahoo! (www.yahoo.com), Infoseek (www.infoseek.com), Lycos (www.lycos.com), NorthernLight (www.nlsearch.com).

Las tareas de un motor de búsqueda en la web son: selección, indexación, búsqueda y visualización de documentos. En la primera tarea, se seleccionan los documentos que serán indexados para su posterior recuperación.

La tarea de indexación de documentos significa construir archivos de acceso a éstos. Esta tarea produce un archivo, que es llamado índice invertido, que se construye a partir del cuerpo de documentos, realizando un mapeo entre cada palabra presente en el cuerpo hacia los documentos que la contienen. La ocurrencia de un término en un documento se llama *posting*. El conjunto de postings asociados a un término se almacena en una *posting list*. Las posting lists también pueden contener otra información como por ejemplo la frecuencia, peso y posición del término en cada documento. La posición de las palabras en el texto se utiliza para realizar búsquedas por adyacencia. El peso se utiliza para un posterior ranking de importancia en función de la cantidad de ocurrencias del término buscado, el lugar de la página donde aparece (en un título por ejemplo) y la forma en la que aparece (enfatizado o no).

Al igual que en la recuperación de información tradicional no se consideran las palabras no significativas o stopwords para la construcción de este índice invertido. Las *stopwords* son palabras que generalmente no describen conceptos y que ocurren frecuentemente en el texto de un documento [Baeza et al., 1999]. Pueden incluir artículos, preposiciones, conjunciones, etc. Estas palabras son removidas durante el análisis del texto de los documentos y de las consultas del usuario, ya que no aportan información significativa a las búsquedas.

La tarea de búsqueda se puede realizar de diversas formas: por una o más palabras, por raíces de palabras o por frases, utilizando para ello, si es necesario, operaciones de lógica de primer orden.

Los algoritmos de búsqueda hacen un ranking de las respuestas según su importancia. Para esto se tiene en cuenta también información de los enlaces, pues un enlace representa una relación entre páginas. Estos enlaces se utilizan para dar peso a una página. La principal diferencia entre los algoritmos de la recuperación de información tradicional y de la recuperación de información en la web es la presencia masiva de estos enlaces o *links*. En la recuperación de información clásica un documento se considera importante si fue citado muchas veces. Una analogía con esto sería considerar una página como importante si hay muchas otras con enlaces a ella.

El motor de búsqueda Google utiliza estas técnicas para realizar un ranking de sus respuestas, utilizando un cálculo de la probabilidad de alcanzar una página dada. El algoritmo que utiliza se denomina PageRank y fue diseñado por integrantes de la Universidad de Stanford [Page et al., 1998]. En este algoritmo, una página tiene un peso alto si la suma de los pesos de sus enlaces entrantes (o in-links) es alta. Un enlace entrante o *in-link* de una página p es un enlace desde una página hacia la página p. Un enlace saliente o *out-link* de una página p es un enlace desde la página p hacia otra página. Entonces, una página con un alto PageRank tiene muchos in-links o pocos in-links con mucho ranking.

Otro algoritmo de ranking de resultados es HITS (Hyperlink Induced Topic Search) [Kleinberg, 1998] [Kleinberg, 1999]. Este algoritmo resuelve el problema de abundancia de documentos dando una medida de la calidad de éstos, distinguiendo entre las páginas, cuáles son las más confiables y centralizadoras. Una página es confiable cuando tiene una gran cantidad de in-links; y una página es centralizadora cuando tiene muchos out-links. La mejor confiabilidad proviene de in-links desde buenas páginas centralizadoras. La mejor centralización proviene de out-links a páginas de buena confiabilidad. El principio general de este algoritmo es calcular un valor de centralización y de confiabilidad de una página a través de la propagación iterativa del peso de confiabilidad y del peso de centralización.

Una diferencia entre PageRank y HITS es que el primero se calcula para todas las páginas web recopiladas por el motor y almacenadas en su base de datos, previo a la realización de consultas. En cambio HITS se ejecuta sobre el conjunto de páginas web recuperadas para cada consulta, en tiempo real. Otra diferencia es que HITS se basa en el cálculo de confiabilidad y centralización, en cambio PageRank se basa sólo en el cálculo de confiabilidad.

Otro tema que es investigado por varias instituciones es la comunicación multilingual en la web y la recuperación de información cross-lingual [Ballesteros, 2001], [Eichmann et al., 1998]. Una introducción al tema se da en algunos estudios del [CLIR]. Muchos motores de búsqueda tienen búsqueda multilingual. Por ejemplo, Open Text Web Index (index.opentext.net) busca en cuatro idiomas: inglés, japonés, español y portugués. Un estado del arte sobre este tema es presentado en [López-Ostenero et al., 2003].

Uno de los problemas que surgen con los motores de búsqueda de la web es que los usuarios no tienen el tiempo y el conocimiento para seleccionar el o los motores más adecuados para su necesidad de información. Una solución posible a esto son los motores de meta búsqueda que son servidores web que envían la consulta a varios motores de búsqueda; recopilan estos resultados y los unifican, uniéndolos y presentándoselos a los usuarios. Algunos ejemplos de meta buscadores son: MetaCrawler (www.metacrawler.com), Dogpile (www.dogpile.com), Copernico (www.copernic.com) y SavvySearch (www.search.com).

Respecto a los usuarios al realizar consultas en la web, las estadísticas [Kobashayi et al., 2000] [Tsikrika, 2001] indican que el número promedio de palabras que utilizan por consulta es de 2 palabras. El número de operadores lógicos por consulta es de 0,4; lo que indica que en general los usuarios no utilizan estos operadores. Las veces que se repiten las consultas es cuatro (en un rango de 1 a 1.5 millones). Por sesión, un usuario hace en promedio dos consultas. El 80% de los usuarios no modifica su consulta inicial y el 85% ve sólo la primera página de la lista de documentos recuperados. Esto indicaría que la gran mayoría de los usuarios desconoce las técnicas de recuperación de información, y tiene dificultad para expresar claramente su necesidad de información, y por lo tanto, no obtienen los resultados deseados. Además, el 85% de los usuarios de Internet utiliza motores de

búsqueda para encontrar información y en general, no están conformes con la performance brindada por estos motores, ni con la calidad de los resultados obtenidos.

Si bien los usuarios no tienen porqué conocer las técnicas de recuperación de información, ya que esto es propio de un especialista en ciencias de la información, se mejorarían los resultados de su búsqueda por medio de una interfase que implemente estas técnicas. En esta tesis, se propone mejorar los resultados de una búsqueda por medio de una estrategia de búsqueda adecuada construida a partir de recursos lingüísticos. Para esto se realiza un refinamiento semántico que le sugiere al usuario la desambiguación de los términos de búsqueda, le permite la selección de un término jerárquicamente relacionado más cercano a su necesidad, y realiza la expansión semántica y multilingual de los conceptos.

2.3. Algunas técnicas utilizadas en la Recuperación de Información

Se han desarrollado numerosas técnicas o herramientas para mejorar la recuperación de información. Una de ellas es el *stemming*. Esta técnica consiste en obtener la raíz de las palabras, de forma que el proceso de búsqueda se realice sobre las raíces y no sobre las palabras originales. El stemming permite a un sistema de recuperación de información relacionar términos presentes en la consulta con los que se encuentren en los documentos y que aparezcan en alguna de sus variantes morfológicas. Para esto se supone que dos palabras que tengan la misma raíz representan el mismo concepto.

Los primeros algoritmos de stemming se desarrollaron para el idioma inglés. Pero esta técnica necesita ser adaptada para lenguas que presentan características distintas al inglés, como ser idiomas más flexivos, tal como el español. Uno de los algoritmos más utilizados para el inglés, es el de Porter [Porter, 1980]. También existen algoritmos para otras lenguas tales como el francés [Savoy, 1999], el español [Figuerola et al., 2002], el holandés [Kraaij & Pohlmann, 1994], el griego [Kalamboukis, 1995] y el latín [Schinke et al., 1996]. En general, estos algoritmos se basan en un conjunto sencillo de reglas que truncan las palabras hasta obtener una

raíz común.

En lenguas aglutinativas, como el alemán y el holandés, en las cuales se unen palabras para formar otras más largas, otra técnica que se puede aplicar es la *segmentación de palabras compuestas* [Monz & de Rijke, 2001]. Por ejemplo, la palabra alemana "Fachinformationszentrum", está compuesta por "Fach" (especialidad), "Information" (información) y "Zentrum" (centro), y se traduce como "centro de información especializada". Diversos estudios muestran que la descomposición de estas palabras en lemas individuales produce una significativa mejora en las búsquedas en este tipo de lenguas, al considerar cada elemento de la palabra compuesta como un término.

Por otro lado, en el entorno de búsqueda tradicional, el usuario debe dividir su interés de búsqueda en distintos conceptos. No siempre un término representa en forma adecuada un concepto. Utilizar otros términos equivalentes o más adecuados para expresar un concepto es realizar una *expansión de consulta* [Efthimiadis, 1996]. Esta situación requiere un cambio en el pensamiento del proceso para elegir los términos de búsqueda. Podría ser necesario consultar recursos lingüísticos, tales como un tesauro o un diccionario, para incorporar nuevos términos. La expansión de consultas es el proceso de suplementar la consulta original con términos adicionales. Durante la expansión también puede haber reemplazo e incluso borrado de los términos correspondientes a la consulta.

2.4. Indicadores de la Recuperación de Información

En la Recuperación de Información se han propuesto diferentes indicadores para medir cuantitativamente la performance de los sistemas de recuperación de información clásicos [Losee, 1998], la mayoría de los cuales pueden ser extendidos para evaluar la búsqueda en la web. Las medidas que se definen en un modelo básico de Recuperación de Información son precisión y recall. La *precisión* se define como el ratio de documentos relevantes sobre el número total de documentos recuperados y el *recall*, también conocido como sensibilidad, se define como la proporción de los documentos relevantes que son recuperados. Es decir:

$$\text{Precisión} = \frac{\text{Número de documentos relevantes recuperados}}{\text{Número de documentos recuperados}}$$

$$\text{Recall} = \frac{\text{Número de documentos relevantes recuperados}}{\text{Número total de documentos relevantes}}$$

Por ejemplo, consideremos una base de datos que contiene 500 documentos y 50 correspondientes a la definición del problema. El sistema recupera 75 documentos, pero solo 45 corresponden a la definición del problema. Los resultados obtenidos de recall y precisión son:

Recall = 45 / 50 = 0.9 es decir el Recall es del 90%

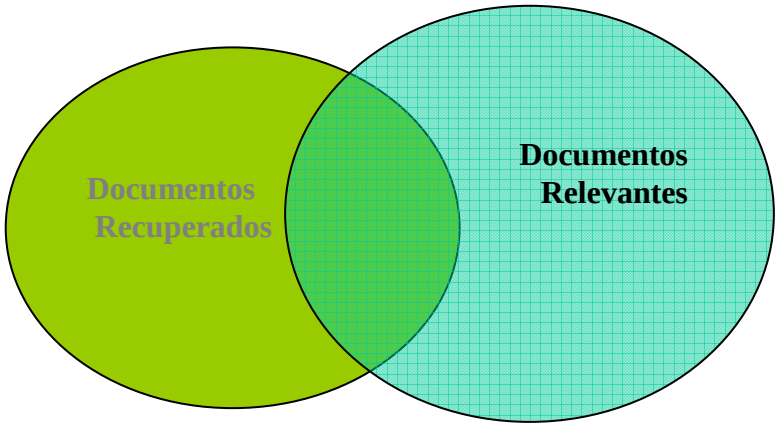
Precisión = 45 / 75 = 0.6 es decir la Precisión es del 60%

Estos indicadores están inversamente relacionados. Es decir, cuando la Precisión aumenta, el Recall normalmente baja y viceversa.

La precisión depende del nivel del usuario: los usuarios experimentados pueden trabajar con un Recall alto y una Precisión baja, porque son capaces de examinar la información y rechazar fácilmente la irrelevante. Los usuarios novatos, por otro lado, necesitan más alta Precisión porque les falta experiencia. Si la tolerancia para errar es alta y la tarea no es en tiempo crítico, esto puede ser aceptable para permitir al usuario revisar varios documentos hasta encontrar si uno es apropiado. De todos modos, si el tiempo es importante y el costo de cometer un error es alto, entonces la Precisión requerida es más alta. Además, cuando los términos son muy específicos aumenta la Precisión y baja el Recall. En cambio, cuando los términos son muy amplios aumenta el Recall y baja la Precisión.

En general, un buen sistema de recuperación de información debe tratar de maximizar la recuperación de documentos relevantes, y minimizar la cantidad de los documentos irrelevantes recuperados.

Gráficamente:



Realizada una búsqueda en una colección de documentos, el conjunto de documentos recuperados no coincide totalmente con el conjunto de los relevantes sobre el tema de interés. Una búsqueda será óptima cuando estos dos conjuntos coincidan, es decir cuando todos los documentos recuperados sean relevantes y todos los documentos relevantes sean recuperados. Estos indicadores, que provienen de la IR tradicional se aplican a la Recuperación de Información en la Web.

2.5. Extracción de Información

La extracción de información o Information Extraction (IE) es una metodología que extrae información, pertinente a las necesidades del usuario, a partir grandes volúmenes de textos. [Bear et al., 1998] proponen utilizar la extracción de información como un post-filtro aplicado a la salida de un sistema de recuperación de información, en vistas a mejorar los resultados de una búsqueda.

El objetivo de un sistema de recuperación de información es consultar una base de datos documental de gran tamaño y devolver un subconjunto de documentos ordenados en forma decreciente según su relevancia respecto al tópico planteado en la consulta. Se considera que tiene éxito si una gran proporción de los documentos devueltos, tomando como base los documentos relevantes existentes en la base de datos, son relevantes de acuerdo al tópico propuesto, y si está correctamente ordenada la respuesta. Es decir, si los documentos más relevantes están situados antes que los menos relevantes.

El objetivo de un sistema de extracción de información es consultar un grupo de documentos, normalmente más pequeño que aquel involucrado en la búsqueda de un sistema de recuperación de información, y extraer ítems pre-especificados de información. Esto puede ser definido especificando instancias de plantillas modelo que deben ser completadas automáticamente sobre la base de un análisis lingüístico de los textos del cuerpo de documentos. Así, se puede afirmar que un sistema tiene una buena performance si el material que extrae captura información relevante.

Desde la perspectiva orientada al usuario [Cuningham, 1999], la IE es un proceso que toma como entrada datos no estructurados y produce como salida datos estructurados. Estos datos pueden ser usados directamente para mostrarse a los usuarios o pueden almacenarse en una base de datos.

Mientras la recuperación de información encuentra documentos y los presenta al usuario, la extracción de información analiza los textos y los presenta sólo si la información específica de ellos es de interés para el usuario. Por ejemplo, un usuario de un sistema de recuperación de información que desea información sobre “estadísticas sobre cáncer de pulmón”, escribirá una lista de las palabras relevantes y recibirá como respuesta un conjunto de documentos, por ejemplo artículos de revistas o noticias de periódicos, que contienen términos coincidentes. Luego, el usuario deberá leer los documentos y extraer él mismo la información que necesita. Como paso siguiente, el usuario podría copiar la información en una planilla de cálculo y producir un gráfico para un reporte. Si este usuario, utilizara un sistema de extracción de información podría, con una aplicación apropiadamente configurada, completar automáticamente la planilla.

Es decir, la recuperación de información recupera documentos relevantes de

las colecciones; mientras que la extracción de información extrae información relevante de los documentos. Así, las dos técnicas son complementarias, y usadas en combinación proveen herramientas poderosas para el procesamiento de texto [Gaizauskas et al., 1998].

La recuperación y la extracción de información también difieren en las técnicas que usualmente aplican. Estas diferencias se apoyan en sus objetivos y en las áreas a partir de las cuales surgieron. La mayoría del trabajo en extracción de información ha surgido de la investigación en sistemas basados en reglas en la lingüística computacional y el procesamiento de lenguaje natural, mientras que la teoría de la información, la teoría de la probabilidad y la estadística han influido en la recuperación de información. Un factor importante para el desarrollo de la extracción de información es el crecimiento exponencial de la cantidad de datos textuales en línea.

Según [De Rosa et al., 2000] existen problemas involucrados con la representación de la información y el proceso de extracción de la información. Para acceder y clasificar la información contenida en los sitios web concernientes a un dominio específico, se necesita representar el dominio, la estructura del sitio y de las páginas web así como la terminología sobre dicho dominio.

En el caso de páginas html, la extracción identifica las instancias de un concepto dado en una página dada, y se analizan en primer lugar los datos no estructurados o textuales. Para esto se utilizan procedimientos de pattern-matching o técnicas de procesamiento de lenguaje natural (Natural Language Processing, NLP) para entender los términos involucrados. Luego se analiza la estructura para encontrar regularidades, como ser tablas o listas, que permitan interpretar los datos.

Como ya se dijo, el objetivo de la extracción de información es transformar texto sin estructura a un formato estructurado. [Eikvil, 1999] diferencia el vínculo entre la extracción de información y texto libre, datos estructurados y datos semiestructurados.

Un texto libre podría ser “nuevos artículos sobre terrorismo”, donde la información clave serían los delincuentes, la ubicación del atentado, la afiliación a la que pertenecen los delincuentes, las víctimas, etc.; o un texto libre podría corresponder a los resúmenes sobre resultados de investigaciones. En el caso del

texto libre, los sistemas de IE utilizan técnicas de lenguaje natural, con reglas de extracción basadas en patrones con análisis sintáctico y semántico y, a pesar que estas técnicas no son comparables a la capacidad humana, proveen resultados útiles.

Respecto a los datos estructurados, éstos se vinculan con la información textual existente en una base de datos. Conocido el formato, la extracción de información requiere técnicas simples y se obtiene un resultado exacto.

Los datos semiestructurados se encuentran en un punto intermedio entre las colecciones no estructuradas (texto libre) y los datos estructurados. Este es el caso de los documentos de la web donde, aunque existe cierta estructuración representada por los tags de html, éstos no son gramaticales; es decir, no indican oraciones completas y no siguen un formato rígido. Entonces, las técnicas desarrolladas para el procesamiento de lenguaje natural no son suficientes por sí solas, como tampoco lo son las reglas simples aplicadas en los datos estructurados.

Dado que la web consiste primariamente de texto semiestructurado, la extracción de información es esencial para cualquier esfuerzo que pretenda utilizar la web como un recurso para el descubrimiento de conocimiento. Un sistema de extracción de información puede pensarse como un intento para convertir información de diferentes documentos de texto en entradas de una base de datos.

Un elemento clave de los sistemas de extracción de información es un conjunto de reglas de extracción del texto o patrones de extracción que identifican la información relevante a extraer [Soderland, 1999].

Según [Cunningham, 1999], hay cinco tareas que debe realizar la extracción y que actualmente se encuentran en investigación y desarrollo, afirmación que coincide con los resultados de las MUC (Message Understanding Conferences). Estas tareas son: Reconocimiento de Entidades Nombradas, Resolución de Co-referencias, Construcción de Elementos Template, Construcción de Relaciones Template, y Producción de Template de Escenarios.

La performance de cada tarea de la extracción, y la facilidad con la cual puede ser desarrollada, varía según el tipo de texto, el dominio o amplitud temática del texto, el estilo en que fueron escritos los textos (informal o formal), y el escenario, es decir, tipos de eventos particulares en los que el usuario de extracción

de información está interesado. Así, una aplicación particular de ésta podría configurarse para procesar artículos (tipo de texto) de noticias financieras (dominio) de un proveedor de noticias particular escritas informalmente (estilo), y encontrar información sobre fusiones de empresas (escenario).

2.6. Sobre Diccionarios, Tesauros y Ontologías

En esta sección se describen recursos lingüísticos tales como diccionarios, diccionarios multilinguales, tesauros y ontologías. Estos recursos se utilizan en el refinamiento semántico propuesto en esta tesis para mejorar la recuperación de información. Se los utiliza para desambiguar los conceptos en el caso de tener varias acepciones, para permitir la selección de conceptos jerárquicamente relacionados y para expandir semántica y multilingualmente cada concepto.

Diccionarios:

Un diccionario indica las distintas acepciones de un término y permite su expansión con sinónimos. Algunos de los diccionarios permiten además la expansión con otros términos relacionados jerárquica y/o semánticamente a cada acepción del término, como ser merónimos, hipónimos e hiperónimos.

La *sinonimia* es la relación entre términos con un mismo significado. Por ejemplo, el término “cáncer” tiene el mismo significado que el término “neoplasma”.

La *meronimia* es la relación semántica entre un término que denota una *parte* y el que denota el correspondiente *todo*. Por ejemplo, el término “brazo” es una parte (merónimo) del término “cuerpo”.

La *hiponimia* es una relación de subordinación entre términos, es decir un término es un hipónimo de otro término si su significado está incluido en el del segundo. Por ejemplo, el término “leucemia” es un tipo (hipónimo) de “cáncer”.

La *hiperonimia* es una relación de superordenación entre términos, es decir un término es un hiperónimo de otro término si su significado incluye al del segundo. Por ejemplo, el término “tumor” (hiperónimo) incluye al término “cáncer”.

Un diccionario muy utilizado como recurso es WordNet [Miller, 1995], el cual puede ser descargado de Internet, o se puede consultar en línea. WordNet es un sistema de referencia léxica online, cuyo diseño está inspirado en teorías psicolingüísticas actuales. Los sustantivos, verbos, adjetivos y adverbios están organizados en conjuntos de sinónimos cada uno de los cuales representa un concepto subyacente. Estos conjuntos de sinónimos además se relacionan jerárquicamente. Este sistema provee las distintas acepciones de un concepto, permitiendo además la expansión de éste con sinónimos, merónimos, hipónimos y otros tipos de términos relacionados a la acepción elegida.

Diccionarios multilinguales:

Para aumentar el número de documentos a recuperar se puede ampliar cada concepto en los idiomas deseados por los usuarios mediante el uso de diccionarios multilinguales generales y especializados disponibles en línea que permiten traducir un concepto a otros idiomas. En el Apéndice 1 se presentan algunos diccionarios multilinguales disponibles en la web.

Tesauros:

La flexibilidad y variedad del lenguaje natural crea serias dificultades para el manejo automatizado de la información. Para solucionar este problema, surgen los tesauros, que permiten el control del vocabulario para representar en forma unívoca cada concepto.

Según la definición de la UNESCO, un tesauro es un instrumento de control terminológico utilizado para traducir a un lenguaje más estricto el idioma natural empleado en los documentos y así asignar palabras claves que describan a cada documento.

Por su estructura, es un vocabulario controlado y dinámico de términos relacionados semántica y genéricamente, los cuales cubren un dominio específico del conocimiento.

El tesauro está estructurado formalmente con el objeto de hacer explícitas las relaciones entre conceptos. Está constituido por términos organizados mediante relaciones entre ellos y provistos de notas de alcance o de definición de los conceptos.

La estructura de la terminología de un tesauro está basada en las interrelaciones entre los conceptos. Estas interrelaciones pueden ser: jerárquicas, de afinidad, y preferenciales. Las relaciones jerárquicas indican términos más amplios o más específicos de cada concepto. Las relaciones de afinidad muestran términos relacionados conceptualmente, pero que no están ni jerárquica ni preferencialmente relacionados. Las relaciones preferenciales se utilizan para indicar cuál es el término preferido o descriptor entre un grupo de sinónimos; y para la calificación de homónimos eligiendo un significado preferido para cada término para diferenciar su significado. En los tesauros las relaciones preferenciales se indican con USE (usar) y SEE (ver), o sus recíprocos UF (Used For, usado por) y SF (Seen For, visto por). Las jerárquicas se representan con BT (Broader Term, término amplio) y NT (Narrower Term, término específico). Las relaciones de afinidad se indican con RT (Related Term, término relacionado).

En el lenguaje natural, existen sinónimos y homónimos. Los sinónimos son grupos de palabras que representan el mismo concepto, por ejemplo agua y H2O. Los homónimos son palabras que representan más de un concepto, por ejemplo banco, que puede referirse al mueble o a la institución financiera. El control de vocabulario implica la selección de un término preferido, también conocido como descriptor o palabra clave, entre un grupo de sinónimos; y la calificación de homónimos eligiendo un significado preferido para cada término [Lancaster, 1995].

En los tesauros se utiliza USE para indicar cuál es el término preferido en el caso de sinónimos. Por ejemplo:

drug addiction
USE substance dependence

Una entrada de esta forma en el tesauro indica que, si se desea encontrar información sobre la adicción a las drogas, la frase *drug addiction* no está permitida porque no es una palabra clave. El tesauro indica que se debe utilizar la frase "substance dependence". Los términos prohibidos se representan en letra cursiva.

UF (Used For) es la relación inversa de USE.

substance dependence
UF *drug addiction*
UF *drug dependence*

En este ejemplo se puede notar que las relaciones inversas también son mostradas en los tesauros. Es decir, que el término "substance dependence", debe utilizarse como término preferido no sólo de *drug addition* sino también de *drug dependence*.

Para homónimos o para indicaciones de múltiples alternativas se utiliza SEE.

processing
SEE fabrication
OR reprocessing

En este caso, el término *processing* es prohibido porque representa más de un concepto. El tesauro indica cuáles serán los términos adecuados para cada significado.

La relación inversa de SEE es SF (Seen For).

fabrication
SF *processing*

Así como UF mostraba la relación inversa para el caso de sinónimos, SF muestra las relaciones inversas para los homónimos.

Para las relaciones jerárquicas se utiliza la sigla BT (Broader Term) para indicar conceptos más amplios.

LUNG NEOPLASMS	LUNG NEOPLASMS
BT1 Respiratory Tract Neoplasms	NT1 Carcinoma Bronchogenic
BT2 Thoracic Neoplasms	NT1 Coin Lesion, Pulmonary
BT3 Neoplasms	NT1 Pancoast’s Syndrome
	NT1 Pulmonary Blastoma

Por ejemplo, el concepto Neoplasms incluye al concepto Lung Neoplasms y éste incluye a Pulmonary Blastoma, que es un tipo particular de cáncer de pulmón.

La relación recíproca se indica con la sigla NT (Narrower Term) para indicar un concepto más específico. Pulmonary Blastoma es un término específico dentro de Lung Neoplasms.

La sigla RT se utiliza para mostrar conceptos relacionados conceptualmente con carácter horizontal. Este tipo de relación se establece entre términos que no son sinónimos ni pueden relacionarse jerárquicamente, pero que permiten una asociación entre ellos; revelando así términos alternativos que hubieran sido útiles en la indización de un documento o en la recuperación de la información.

AGE GROUPS	ADULTS
RT Adolescents	RT Age groups
RT Adults	
RT Children	
RT Infants	

En el ejemplo se observa que los grupos etarios (AGE GROUPS) tiene como términos relacionados Adolescentes, Adultos, etc.

En las bases de datos documentales se utilizan palabras claves para describir el contenido de un documento. Estas palabras claves, o descriptores, pueden estar formadas por un término o por una frase que se eligen de un diccionario de términos controlados o permitidos para el sistema, es decir, de un tesauro. Así, el tesauro representa una herramienta documental que permite la conversión del lenguaje natural de un documento al lenguaje controlado documental [Lancaster, 1995].

Los términos del tesauro se clasifican en *descriptores*, o términos principales o preferidos o permitidos; y *no descriptores*, es decir, términos equivalentes de carácter secundario o no preferidos o prohibidos. Los términos no descriptores no pueden ser utilizados como palabras claves de los documentos para su indización ni como términos de búsqueda. Para cada término no descriptor, el tesauro indica cual es el término permitido correspondiente para representar dicho concepto.

A diferencia de un diccionario, donde todos los sinónimos de un concepto son representativos y tratados por igual, en un tesauro se tiene una palabra clave preferida y representativa del conjunto de sinónimos para cada concepto.

La mayoría de los tesauros existentes están actualmente disponibles en línea. Como un aporte de esta tesis, se realizó un relevamiento de tesauros disponibles en línea. En el Apéndice 1 se presentan estos tesauros agrupados según el área del conocimiento con sus respectivas direcciones de Internet.

Ontologías:

Las ontologías proporcionan una vía para representar el conocimiento y son un enfoque importante para capturar semántica. La definición más consolidada es la que la describe como "una especificación explícita y formal sobre una conceptualización compartida" ([Gruber, 1993], [Studer, 1998]). Es decir, las ontologías definen conceptos y relaciones de algún dominio, de forma compartida y consensuada; y esta conceptualización debe ser representada de una manera formal, legible y utilizable por las computadoras. Las ontologías consisten de términos organizados en una taxonomía, sus definiciones y axiomas que los relacionan con otros términos.

Tim Berners-Lee, uno de los pioneros de la web semántica, promueve el desarrollo de la web con conocimientos [Berners-Lee, 2001], y organizaciones como SematicWeb [SemanticWeb] se encargan de estandarizar lenguajes y herramientas para dar semántica a la web. La importancia de las ontologías en la web se aprecia con la aparición de agentes de búsqueda de información, que explotarán el conocimiento anotado en las páginas web, serán capaces de interpretar los esquemas ontológicos y axiomas de diferentes dominios, mantendrán la consistencia de las instancias que se inserten en las páginas web siguiendo los esquemas ontológicos definidos y realizarán una búsqueda con inferencias utilizando los axiomas.

Actualmente, los buscadores realizan la búsqueda de información en el texto de las páginas web escritas en código html. Existe, sin embargo, la tendencia de implementar el uso de metadatos para agregar datos sobre los datos. Esto se efectúa mediante anotaciones de datos introducidas dentro del código html, siguiendo algún

esquema de anotación común, normalmente basado en el estándar de intercambio de datos XML.

La idea es que los datos puedan ser utilizados y “comprendidos” por las computadoras sin necesidad de supervisión humana, de forma que los agentes web puedan ser diseñados para tratar la información situada en las páginas web de manera semiautomática. Es decir, convertir la información en conocimiento, referenciando datos dentro de las páginas web a metadatos con un esquema común consensuado sobre algún dominio. Los metadatos no sólo especifican el esquema de datos que debe aparecer en cada instancia, sino que además pueden tener información adicional de cómo hacer deducciones con ellos, es decir, axiomas que podrán aplicarse en los diferentes dominios que trate el conocimiento almacenado. Con ello, se mejora la búsqueda de información, ya que las anotaciones seguirán un esquema común que podrá ser aprovechado por los buscadores web.

Los agentes de búsqueda en la web no sólo encontrarán la información de forma precisa, si no que podrán realizar inferencias automáticamente buscando información relacionada con la que se encuentra situada en las páginas, y con los requerimientos de la consulta indicada por el usuario.

Las ontologías tienen los siguientes componentes que sirven para representar el conocimiento sobre un dominio:

Conceptos: son las ideas básicas que se intentan formalizar. Los conceptos pueden ser clases de objetos, métodos, planes, estrategias, procesos de razonamiento, etc.

Relaciones: representan la interacción y enlace entre los conceptos del dominio. Suelen formar la taxonomía del dominio. Por ejemplo: subclase-de, parte-de, conectado-a, etc.

Funciones: son un tipo concreto de relación donde se identifica un elemento mediante un cálculo que considera varios elementos de la ontología. Por ejemplo, pueden aparecer funciones como categorizar-clase, asignarfecha, etc.

Instancias: se utilizan para representar elementos o individuales determinados de un concepto.

Axiomas: son teoremas que se declaran sobre relaciones que deben cumplir los elementos de la ontología. Por ejemplo: “Si A y B son de la clase C, entonces A no es subclase de B”, “Para todo A que cumpla la condición C1, A es B”, etc. Los axiomas, permiten junto con la herencia de conceptos, inferir conocimiento que no esté indicado explícitamente en la taxonomía de conceptos. Los axiomas sirven para modelar predicados que son siempre verdaderos. Los axiomas comúnmente se utilizan para representar conocimiento que no puede ser formalmente definido por otros componentes, para verificar la consistencia de la ontología misma, y durante el proceso de inferencia de nuevo conocimiento.

Para poder explotar la web semántica, se necesitan lenguajes de marcado apropiados que representen el conocimiento de las ontologías. El lenguaje XML con sus respectivos DTD (Document Type Definition) no es suficiente para esto. [Decker et al. 2000] [Broekstra et al. 2002]. Existen otros lenguajes de marcado como ser RDF (Resource Description Framework), recomendado por el consorcio W3C como estándar para los metadatos. Mediante anotaciones RDF y RDF Schema se pueden representar algunos aspectos sobre conceptos de un dominio y, mediante relaciones taxonómicas, crear una jerarquía de conceptos. Existen herramientas disponibles como Protégé[3], OntoEdit[4], y WebOnto[5] para realizar anotaciones en documentos con lenguajes de marcado propios. Un lenguaje con gran capacidad expresiva que ha emergido como un estándar para realizar anotaciones de ontologías en la web es OWL (Ontology Web Language) [OWL]. OWL es un lenguaje de marcado semántico para publicar y compartir ontologías en la web y es el lenguaje de ontologías para la web, desarrollado por el W3C. OWL está dividido en tres tipos (OWL Lite, OWL DL y OWL Full), cada una de los cuales provee diferentes niveles de expresividad. OWL Lite permite la construcción de una taxonomía e incluye la posibilidad de expresar igualdades, desigualdades y restricciones simples. OWL DL permite la máxima expresividad, manteniendo a su vez la completitud computacional (todas las conclusiones se garantizan que son computables) y decibilidad (todos los cómputos finalizarán en un tiempo finito). OWL Full tiene la máxima expresividad,

[3] protege.semanticweb.org
[4] ontoserver.aifb.unikarlsruhe.de/ontoedit/
[5] kmi.open.ac.uk/projects/webonto/

libertad sintáctica, y compatibilidad full con RDF Schema, aunque se pierden las garantías computacionales.

Al utilizar una ontología en conjunción con un razonador o motor de inferencias, éste puede realizar deducciones en torno a dicho modelo. Entre las funciones más comunes que pueden ofrecer dichos razonadores se encuentran el chequeo de inconsistencias, la clasificación de instancias y conceptos, la inferencia de relaciones inversas, simétricas y transitivas, y el razonamiento con reglas de inferencia dadas por el usuario.

Para potenciar el uso de ontologías en la web, se necesitan aplicaciones específicas de búsqueda de ontologías, como OntoAgent[6] que indiquen a los usuarios las ontologías existentes y sus características para poder utilizarlas en su sistema, y como OntoSeek [Guarino et al., 1999] para la búsqueda de información.

A diferencia de los tesauros y de los diccionarios, en las ontologías, además de representar las relaciones entre conceptos, se agregan los axiomas, que permiten realizar inferencias sobre los conceptos.

2.7. Utilización de los recursos lingüísticos

En la Sección 2.6 se han descrito recursos lingüísticos que se utilizan como ayuda para la preparación de estrategias de búsqueda adecuadas que representen la necesidad de información del usuario.

En el problema presentado en el Capítulo 1, se mostró que el usuario puede recurrir a distintas fuentes para recuperar información, como ser bases de datos documentales y páginas web, cada una de ellas con características propias.

En las bases de datos documentales, los documentos son recopilados y analizados por instituciones especializadas que asignan palabras claves o términos controlados a los documentos, utilizando un tesauro. Para la búsqueda de información en estas bases de datos el uso de tesauros permite obtener un resultado más preciso. Esto se debe a que en el caso de sinónimos el tesauro indica cuál es el

[6] delicias.dia.fi.upm.es/OntoAgent

término preferido que se utiliza como descriptor en los documentos. La utilización de términos preferidos aumenta la precisión en la búsqueda. Si se desea aumentar la cantidad de documentos a recuperar puede utilizarse también en la consulta los sinónimos del término preferido, resignando la precisión. Por otro lado, la estructura jerárquica de los tesauros permite que un usuario pueda seleccionar un concepto más específico a su interés de búsqueda, y de este modo mejorar la precisión de los resultados.

En las páginas web la terminología no está controlada. Es decir, en la web no existe una representación unívoca de los conceptos y distintos autores pueden utilizar términos distintos para referirse a un mismo concepto.

Una forma de resolver este problema es incorporar a la búsqueda los sinónimos de cada concepto a buscar, aumentando así la cantidad de documentos a recuperar. Esto se puede realizar utilizando diccionarios. A diferencia de un tesauro, donde se tiene una palabra clave preferida y representativa del conjunto de sinónimos para cada concepto, en un diccionario todos los sinónimos de un concepto son representativos y tratados por igual. Para la expansión semántica, un tesauro puede ser usado como diccionario si para cada conjunto de sinónimos se ignora el término preferido y se trata a todos los sinónimos por igual. Otra posibilidad es la utilización de ontologías, que a diferencia de los tesauros y de los diccionarios, en las ontologías, además de representar las relaciones entre conceptos, agregan los axiomas, que permiten realizar inferencias sobre los conceptos. Los conceptos obtenidos a partir de las inferencias y de las relaciones en una ontología se utilizan también en la expansión de la consulta [Deco et al., 2005c].

Por esto, el uso de tesauros es más adecuado en el caso de *búsqueda en bases de datos*, y el uso de diccionarios y de ontologías para la *búsqueda de información en páginas web* ya que en este caso la terminología no está controlada.

En el Capítulo 4 se presenta la arquitectura de un refinador semántico que, a partir de los conceptos ingresados por el usuario, construye una estrategia de búsqueda que represente su necesidad de información, utilizando estos recursos lingüísticos según lo discutido en esta sección.

Los recursos lingüísticos no sólo pueden utilizarse para la preparación de la estrategia de búsqueda, sino también para la clasificación e integración de la información.

En la *clasificación* de la información resultante de las páginas web obtenidas a través de un buscador, estos recursos permiten reconocer conceptos similares. Tener, por ejemplo, una ontología que agrupe los conceptos {'proyecto de investigación', 'trabajo de investigación', 'research project'}, ayudaría a clasificar en un mismo grupo, páginas que contengan datos de proyectos de investigación que se estén realizando sobre el tema buscado. He colaborado en el desarrollo de esta idea en las siguientes publicaciones: [Motz et al., 2003a], [Bender et al., 2004], [Bender et al., 2005].

En la *integración* de la información obtenida a partir de las distintas fuentes los recursos permiten unificar conceptos expresados con distinta terminología y reconocer coincidencias de autores o instituciones que puedan estar expresadas de distinta manera. Por ejemplo, reconocer que dos documentos provienen de una misma institución si en los respectivos documentos XML el tag o campo <Institución>, contiene el valor "MIT" en uno de ellos y el valor "Massachusetts Institute of Technology" en el otro. En [Motz et al., 2000] se describe el uso de esta técnica para instanciar bases de datos desde páginas web. En [Motz et al., 2001] se presenta un mecanismo para integrar bases de datos con información extraída de la web.

Capítulo 3: Trabajos relacionados

En este capítulo se comentan trabajos y proyectos relacionados con la recuperación de información en la web.

3.1. Utilización de WordNet

A partir de la existencia de este recurso lingüístico se han realizado muchas experiencias para su aprovechamiento en distintas áreas, entre ellas la Recuperación de Información. A diferencia del Procesamiento de Lenguaje Natural, donde una de las aplicaciones de WordNet[7] es la desambiguación automática del sentido de una palabra, en la Recuperación de Información, WordNet es utilizado para expandir la consulta.

En esta tesis, WordNet se ha utilizado como recurso lingüístico para la preparación de la estrategia de búsqueda. WordNet se usa para mostrarle al usuario los distintos significados de un concepto, sugerirle términos jerárquicamente relacionados con el concepto de su interés e incorporar sinónimos a cada concepto de búsqueda.

[Voorhees, 1998] argumenta que las expansiones con recursos lingüísticos tales como WordNet, son efectivas para consultas con muy pocos conceptos, mientras que no trae mucha mejora para consultas con muchos conceptos.

Sin embargo, [Mandala et al., 1998] concluyen que las expansiones de la consulta con Wordnet pueden mejorar la cantidad de documentos a recuperar pero decrece la precisión. Este decrecimiento de la precisión se debe, según estos autores, a que existen muchas relaciones entre términos que no se encuentran en WordNet y a que hay términos que no están en este recurso, como ser nombres propios. Además, otro problema que señalan es que en la realidad existen términos que están relacionados, como ser “stochastic” y “statistic”, pero que en WordNet no se pueden relacionar porque pertenecen a grupos distintos de este recurso: el primero es adjetivo y el segundo es sustantivo.

[Martínez et al., 2002] presentan un sistema donde la consulta se ingresa en lenguaje natural. De esta frase en lenguaje natural, el sistema primero detecta las palabras de interés para la búsqueda. Luego utiliza recursos lingüísticos para expandir la consulta. Utiliza el recurso Aries[8] para buscar variantes morfológicas de las palabras y el recurso EuroWordNet para buscar sinónimos de las palabras a buscar. Aries es un léxico morfológico para el castellano, desarrollado por la Universidad Politécnica de Madrid, y es un recurso que requiere licencia de uso. EuroWordNet[9] es similar al recurso WordNet pero incluye vocabulario en inglés, español, alemán e italiano; también requiere licencia de uso.

En [Gonzalo et al, 1998] se realizan experimentos sobre documentos indexados en la forma clásica y sobre documentos indexados con los synsets (conjuntos de sinónimos) de WordNet, luego de la desambiguación manual de los términos de los documentos. Entre los resultados obtenidos, se demuestra que si se puede desambiguar la consulta mediante los synsets de WordNet se mejora la performance aunque no se desambiguen los documentos. Además, se muestra que no se degrada la performance, si hay menos de 10 por ciento de errores en la desambiguación de sentido de la palabra.

[Navigli et al., 2002] proponen una adaptación automática de WordNet a distintas áreas del conocimiento. Presentan un método para enriquecer WordNet automáticamente con subárboles de conceptos de un área del conocimiento. En [Navigli et al., 2003] se experimenta la posibilidad de usar información ontológica para extraer el dominio semántico de una palabra. Estos autores proponen la expansión de la consulta considerando las palabras en un “sense definition”, en lugar de utilizar relaciones taxonómicas, como ser sinónimos e hiperónimos. Señalan que los métodos más exitosos de expansión de la consulta parecen sugerir que la mejor forma de expandirla es agregando palabras que a menudo co-ocurren con las palabras de la consulta. Por ejemplo, palabras que, sobre una plataforma probabilística, se cree que pertenecen al mismo dominio semántico, como ser: cáncer y medicina. Los autores presentan un método de desambiguación del sentido de una

[7] www.cogsci.princeton.edu/~wn/
[8] www.mat.upm.es/~aries/
[9] www.let.uva.nl/~ewn/

palabra basado en reconocimiento de patrones estructurados, y usan este método para explorar varias estrategias basadas en sentido para expandir la consulta.

WordNet no es el único recurso lingüístico utilizado para las expansiones de la consulta. Al igual que la propuesta de esta tesis, donde también se proponen utilizar otros recursos tales como tesauros, [Sangoi Pizzato et al., 2003] expanden la consulta utilizando tesauros y muestran que esta propuesta mejora la recuperación de la información en la web.

En [Carpineto et al, 2002] se propone extraer los términos para la expansión de la consulta desde un conjunto inicial de documentos recuperados. Es decir, proponen realizar un feedback de relevancia, incorporando a la consulta palabras de los documentos que el usuario marcó como relevantes para su interés. Por otra parte, [Cui et al, 2000] proponen expandir la consulta a partir de los query logs (logs de consulta) de los usuarios. Es decir, expanden la consulta con términos obtenidos del historial de consultas de dicho usuario.

[Magnini y Cavaglia, 2000] presentan un trabajo cuya hipótesis es que, la inclusión de expansiones léxicas debería traer una mejora en la recuperación de documentos relevantes. La modalidad de expansión que proponen es expandir primero cada palabra clave con sus derivaciones morfológicas y sinónimos, y luego construir una expresión booleana.

Se describen a continuación otros proyectos relacionados con la recuperación de información en la web. La presentación se hace en orden alfabético por el nombre de proyectos. Se presenta luego un cuadro comparativo de los mismos.

3.2. CiteSeer

Algunas publicaciones científicas están disponibles en la web en formato html, lo que permite que el texto de estas publicaciones sea recuperable con los motores de búsqueda de la web. Pero muchos de los documentos publicados en la

web están en formato postscript o pdf, y no en html. Esto presentaba un problema porque el texto de estos documentos no es indexable por los motores de búsqueda.

CiteSeer [Bollacker et al.1998], formalmente llamado ResearchIndex, es un agente de software que asiste al usuario en la búsqueda de publicaciones en la Web. Es un proyecto del NEC Research Institute que mejora el proceso de búsqueda manual de este tipo de documentos porque automatiza el proceso tedioso, repetitivo y lento de encontrar y recuperar publicaciones en la web. Una vez que los potenciales documentos relevantes son recuperados, guía al usuario, sugiriéndole otros documentos relacionados. Para esto usa medidas de similaridad derivadas de características semánticas de los documentos relevantes.

La síntesis de su funcionamiento es la siguiente: dado un conjunto de palabras claves, usa motores de búsqueda web para localizar y descargar documentos potencialmente relevantes al tema buscado por el usuario. Los documentos descargados son parseados para extraer características semánticas, incluyendo información de frecuencia de citación y de palabras. La información se almacena en una base de datos, en la cual el usuario puede buscar por palabras claves o usar enlaces basados en citaciones para encontrar documentos relevantes.

CiteSeer crea automáticamente una base de datos local que estructura los documentos descargables de la web, y permite la búsqueda dentro de documentos en formato postscript o pdf. Esta búsqueda sobre estos formatos de documentos no es realizada por muchos motores de búsqueda. Además, para que un documento sea ingresado en esta base de datos, no se requiere ningún esfuerzo extra por parte de los autores del trabajo dado que el proceso de extracción y carga es automático.

La arquitectura del agente tiene tres componentes principales: un subagente para localizar y adquirir automáticamente publicaciones científicas, un parseador de documentos que crea y carga la base de datos, y una interfaz de navegación para la base de datos que soporta la búsqueda por palabras claves y la navegación por enlaces de citación.

Cuando el usuario desea explorar un nuevo tema, se crea una nueva instancia del agente para ese tema particular. Se invoca a un subagente para buscar páginas web que probablemente contengan documentos de investigación de interés, en formatos postscript o pdf. Para ello el subagente utiliza motores de búsqueda y

heurística, como por ejemplo páginas que contengan las palabras “publication” o “postscript”. Luego, el subagente descarga los archivos, identificándolos por las extensiones .ps, .ps.Z o .ps.gz, y evita descargar archivos duplicados.

El parseado de documentos consiste en procesar los documentos descargados para extraer las características semánticas de éstos. Los programas de parsing extraen los datos de interés de los documentos y los colocan en una base de datos relacional.

La base de datos contiene las siguientes tablas: document, documentwords, citations, citationwords, citecluster y clusterweights. La tabla document contiene piezas de texto del documento, URL del documento, y un único id de artículo. Documentwords contiene información de la frecuencia de palabra sobre el cuerpo del documento referenciado en la tabla document. La tabla citation contiene el texto de las citaciones hechas por el documento en la tabla document, tiene un único id de citación y el id de artículo correspondiente. Citationwords contiene la frecuencia de palabras sobre las citaciones en la tabla citation. Citecluster y clusterweights contienen el número de cluster e información de peso para cuando se agrupan citaciones similares de diferentes formas (esta información es usada para la recuperación de documentos similares).

El subagente extrae el texto ASCII del archivo que contiene el documento, formateado usando información del formato original postscript o pdf. Luego, verifica que este texto ASCII sea un documento de investigación, incluyendo un chequeo de la existencia de referencias o citas al final del documento. Se usa heurística para identificar, en un documento válido, el Header (que es la información al principio del documento que contiene título, autor, institución, etc), el Abstract (que se extrae del mismo), la Introducción (si existe, se extraen las primeras 300 palabras), Citaciones (se extrae la lista de referencias) y Frecuencia de palabras (se graban para todas las palabras excepto para las de las citaciones y las stopwords). Se implementa stemming usando el algoritmo de Porter.

CiteSeer es un Autonomous Citation Indexing (ACI). Un ACI automatiza totalmente el proceso de crear un índice de citaciones, es decir referencias bibliográficas o citas, para literatura en formato electrónico. Luego de parsear el documento uno de los principales problemas que debe resolver un ACI es el de determinar cuando dos citas hacen referencia a un mismo documento.

A cada documento se le extraen de sus referencias bibliográficas: título, autor, año de publicación, número de páginas y la etiqueta de citación. Se usa la etiqueta de citación para encontrar la ubicación en el documento de la cita, lo que permite extraer el contexto de la cita durante un browsing a la base de datos.

El navegador de base de datos consiste en un subagente de procesamiento de consulta que toma la consulta del usuario y retorna una respuesta en formato html, a través de un navegador web. Se pueden realizar búsquedas por palabras clave, ya sea sobre el texto de los documentos o sobre las citas bibliográficas. Después de una búsqueda inicial por palabras claves se puede navegar por los documentos siguiendo las citas como enlaces. Los resultados pueden ser ordenados por cantidad de citaciones, por fecha de publicación, etc.

Para las medidas de distancia semántica se implementa un método de agrupamiento de citaciones idénticas (ICG). El primer paso en este método es una normalización de citaciones con reglas como la conversión a minúsculas y eliminación de puntuaciones. Luego se usa un algoritmo de correspondencia de palabra/frase para agrupar las citaciones. En este algoritmo, si una citación bajo consideración está lo suficientemente cercana a un grupo de citaciones existentes, entonces se la incluye en el mismo. Si no, se crea un grupo nuevo.

Ahora, dada una base de datos de documentos un usuario podría querer encontrar un documento de interés y luego querer encontrar otro documento relacionado. Para ubicar documentos similares usa un mecanismo para la recuperación automática de documentos relacionados basado en la medición de distancia de las características semánticas de éstos. Los agentes asistentes web anteriores han usado información de la frecuencia de palabras para medir automáticamente cuán relacionados están dos documentos.

Las citaciones de otros trabajos que eligen los autores en sus documentos son una buena información para juzgar la relación entre documentos. Se utilizan las citaciones en común para estimar qué documentos en la base de datos están más relacionados al elegido por el usuario. Esta medida se llama Common Citation x Inverse Document Frecuency (CCIDF). CiteSeer también combina los diferentes métodos para lograr una medida de distancia que sea más precisa que un método por sí solo.

Este proyecto está activo y el motor de búsqueda está disponible en http://citeseer.org/.

3.3. InfoSleuth

InfoSleuth[10] [Nodine et al. 2000] [Fowler et al. 1999] es un sistema basado en agentes, diseñado para integrar fuentes y herramientas heterogéneas y distribuidas mediante el uso de ontologías. Un conjunto de agentes de InfoSleuth colabora en el nivel semántico para ejecutar la recolección de información y en las tareas de análisis, donde las fuentes de información subyacentes pueden tener diversas estructuras y contenidos. Las ontologías por sí mismas son vocabularios estructurados que representan la metadata de un dominio de aplicación particular. Es un proyecto desarrollado por la MCC (Microelectronics and Computer Technology Corporation) de Austin Texas.

Una aplicación InfoSleuth es una colección de agentes, codificados en Java para portabilidad y compatibilidad con los web browsers populares. Los agentes se comunican a través del lenguaje Knowledge Query Manipulation Language (KQML), lo que implica comunicación a nivel semántico sobre ontologías. KQML es un lenguaje diseñado para soportar interacciones entre agentes de software inteligentes.

Los agentes utilizan el lenguaje estándar Open Knowledge Base Connectivity (OKBC) para comunicar información sobre sus ontologías y las restricciones en los conceptos en sus ontologías. OKBC es un protocolo que provee un conjunto de operaciones para una interfaz genérica para sistemas de representación de conocimiento subyacente.

La arquitectura de InfoSleuth es dinámica y está basada en agentes. Cada agente provee un conjunto de servicios que se pueden describir como un conjunto de tareas sobre el dominio de interacción. El UserAgent asiste al usuario en las consultas utilizando ontologías y le muestra los resultados. El BrokerAgent hace corresponder las solicitudes de servicios o información con los agentes que pueden

[10] www.argreenhouse.com/InfoSleuth/

proveerlos. El OntologyAgent provee el conocimiento y responde consultas sobre las ontologías. Los ResourcesAgents asocian las consultas con los datos almacenados y los repositorios externos o propios que los contienen.

Un agente de consulta descompone la consulta y la distribuye entre subagentes que acceden a distintos recursos, y luego recompone estos resultados parciales. Hay también otros agentes que realizan funciones especiales, como agregación de datos y detección de eventos.

Los agentes se comunican y razonan sobre la capacidad de los otros agentes en términos de un modelo ontológico de manejo de información para resolver la solicitud del usuario, el cual no necesita conocer nada acerca de la ubicación física o características estructurales de cualquier recurso. Las solicitudes son expuestas en términos de una ontología, llamada la ontología de dominio de la aplicación, que provee una infraestructura semántica para actividades de información en el dominio de interés del usuario. El crecimiento semántico de las comunidades de agentes es soportado denotando la intermediación semántica, mediante brokers, lo que permite a los agentes identificar potenciales colaboradores.

3.4. OntoBroker

Es un sistema de búsqueda basado en ontologías con axiomas. OntoBroker[11] [Decker et al. 1999] [Fensel et al., 1998] utiliza las ontologías para describir páginas web, formular consultas y derivar respuestas. Aplica técnicas de inteligencia artificial para mejorar el acceso a fuentes de información heterogéneas, distribuidas y semiestructuradas. OntoBroker usa lógica Frame Logic para definir la ontología y representar una base del conocimiento que permita la inferencia. La extracción de metadatos de una página web se hace por wrappers o web crawlers que identifican la semántica especial etiquetada en las páginas web.

La arquitectura está formada por un web crawler, una interfaz de consulta y un motor de inferencia. El crawler se encarga de recolectar páginas web, extraer las descripciones semánticas y parsearlas al formato interno de OntoBroker. La

[11] ontobroker.semanticweb.org

información recolectada se almacena en una base de datos. Las descripciones semánticas deben estar hechas en html-A que es una extensión de html definida para este proyecto. Html-A no agrega información a las páginas sino que sólo hace explícita la semántica de los datos ya presentes. Esta tarea de agregar descripciones semánticas es manual, lo cual fue uno de los mayores problemas de OntoBroker.

La interfaz de consulta se utiliza para que el usuario complete campos de un formulario. Se usa un browser de ontologías para encontrar los campos buscados en la ontología. El motor de inferencia utiliza los datos ingresados por el usuario junto a los de la ontología y deduce las respuestas.

Dos problemas significativos que presenta OntoBroker son la lentitud del motor de inferencias para grandes cantidades de datos, y el gran esfuerzo humano para agregar semántica a los documentos html.

On2broker [Fensel et al., 1999] [Fensel et al., 2000] es el sistema sucesor de OntoBroker y resuelve estos problemas. Las nuevas decisiones de diseño de On2broker son la clara separación de consulta y motores de inferencia, y la integración de nuevos estándares web como XML y RDF.

La arquitectura de On2broker está formada por un agente de información, un agente de inferencia y un motor de consulta.

El agente de información recolecta información de la web y soporta lenguajes estándares de descripción de contenido, además del html-A que era propietario. Este agente también utiliza wrappers para extraer información semántica automáticamente.

El agente de inferencia utiliza información de la base de datos y de las ontologías para derivar conocimiento implícito y lo guarda en forma explícita.

El motor de consulta resuelve las consultas usando los contenidos de la base de datos que es relacional.

On2broker ha sido usado en varias aplicaciones. La más prominente es la iniciativa que proporciona el acceso semántico a todos los tipos de información de los grupos de la comunidad de adquisición de conocimiento. Usa información semántica para guiar el proceso de respuesta a una consulta y proporciona las respuestas con una sintaxis y una semántica bien definidas que pueden entenderse

directamente y procesarse por agentes automáticos u otras herramientas de software.

Actualmente, Ontobroker ha madurado y es un software comercial disponible en Ontoprise (http://www.ontoprise.de/) en su versión 5.1.

3.5. OntoSeek

OntoSeek [Guarino et al. 1999] es un sistema diseñado para la recuperación de información desde páginas amarillas y catálogos de productos. Es un ejemplo concreto del uso de ontologías para la recuperación de información y combina un mecanismo de correspondencia de contenido conducido por ontología con un formalismo de representación expresivo.

Es un proyecto de cooperación entre el Consorcio di Ricerca Nazionale Tecnologia Ogeetti (CORINTO) y el National Research Council-Institute of System Science and Biomedical Engineering, que son parte del proyecto de recuperación y reuso de componentes de software orientado a objetos. El proyecto adoptó Java como tecnología para desarrollar una poderosa interfaz de usuario integrada para la web.

OntoSeek tiene asistencia interactiva en la formulación de la consulta, los factores de recall y precisión son buenos, y es eficiente en grandes volúmenes de datos.

El sistema utiliza ontologías. Cuando se planteó el proyecto se decidió evitar construir una ontología de la nada. Se eligió la ontología Sensus[12], la cual consta de cerca de 90.000 nodos, en su mayor parte resultado de combinar tesauros. Sensus es una ontología muy amplia dotada con poderosas interfaces léxicas derivada de WordNet, la cual devuelve la categoría léxica y un sentido asociado a cada palabra.

En la etapa de codificación, el sistema codifica un recurso, que puede ser tanto un documento como un servicio web, descrito en lenguaje natural en un grafo simple de conceptos y relaciones. Para ello emplea grafos conceptuales léxicos (LCG). Los nodos y los arcos etiquetados usados son reconocidos por la interfaz léxica, la cual pregunta para elegir entre cada significado asociado a la palabra, según

[12] mozart.isi.edu:8003/sensus2/

la información en el vocabulario. El grafo de las palabras es por lo tanto traducido dentro de un grafo de significado, cada uno correspondiendo a un nodo en la ontología. Después de la validación semántica, ejecutada con la ayuda de la ontología, la clasificación almacena el LCG en la base de datos.

En el proceso de recuperación de información, el usuario representa la consulta nuevamente como un LCG. Este grafo se somete a desambiguación léxica y validación semántica. El sistema busca en la base de datos los ítems de información descritos por ese grafo. OntoSeek luego presenta las respuestas al usuario como un informe html.

La arquitectura de OntoSeek implementa el típico paradigma cliente servidor. La arquitectura central es un servidor de ontología. El servidor provee una interfaz para aplicaciones que acceden o manipulan un modelo de datos ontológico, y facilidades para mantener una base de datos LCG persistente. Los codificadores de recursos y los usuarios finales pueden acceder al servidor a través de los protocolos de comunicación pregunta/respuesta. La base de datos LCG puede ser también actualizada offline por compiladores, que aceptan como entrada LCGs codificados en lenguajes de marcado, tales como extensiones html o XML.

3.6. WebFind

WebFind [Monge et al. 1996] es una herramienta que descubre documentos científicos que están disponibles en los sitios de sus autores en la web. Es un proyecto de la Universidad de California, San Diego (UCSD).

Usa una combinación de fuentes de información externas como una guía para localizar dónde buscar por información en la web. Estas fuentes son: Melvyn y NetFind. Melvyn es el catálogo online de bibliotecas de la Universidad de California, e incluye bases de datos de registros bibliográficos tales como la base de datos Inspec de ciencia e ingeniería. NetFind es un servicio para encontrar direcciones de email y direcciones de hosts de Internet.

Para recuperar un documento científico en la web, WebFind primero integra la información provista por Melvyn y por NetFind. La búsqueda comienza cuando el

usuario provee palabras claves para identificar el documento. Un documento puede ser identificado usando cualquier combinación de nombres de sus autores, palabras del resumen, u otra información bibliográfica. Una vez que el usuario confirma que se ha encontrado el documento correcto, consulta en las bases de datos de Melvyn para encontrar la asociación institucional del autor principal del documento. Luego usa NetFind para obtener la dirección de Internet de un host con la misma asociación institucional. La consulta a NetFind consiste en un conjunto de palabras claves que describen la institución. En general, en el resultado se obtienen varios hosts para cada institución. Para esto, WebFind usa un algoritmo para hacer un ranking con las direcciones de los hosts para elegir cuál es el mejor.

La búsqueda realizada por WebFind es en tiempo real. La información recolectada de un documento recuperado se analiza y utiliza para decidir qué documentos son recuperados después.

Primero, se trata de encontrar un servidor web en el host de Internet elegido. WebFind usa heurística basada en patrones comunes para nombrar servidores (www. o www-). Prueba la existencia de un servidor usando ping. Si no encuentra ninguno de los prefijos, elimina el primer segmento del nombre de dominio del host y aplica otra vez la misma heurística. En segundo lugar, sigue enlaces hasta que el artículo requerido es encontrado.

La búsqueda procede en dos etapas: encontrar una página web del autor principal y encontrar una página web que sea el artículo deseado. En la primera etapa, el conjunto primario de claves es el nombre del autor principal, y el secundario es: personal, gente, autoridad, etc. Intuitivamente, el objetivo principal es encontrar la página principal del autor y si no la encuentra, localizar una lista de personal en la institución. En la segunda etapa, el conjunto primario de claves es el título del artículo requerido y el secundario es: publicaciones, documentos, reportes, etc. El objetivo principal es encontrar el documento requerido y si no lo encuentra, localizar una página con punteros a documentos en general.

En cada paso, el procedimiento de búsqueda es quitar repetidamente el primer enlace de una cola de prioridad, y recuperar la página apuntada. La búsqueda tiene éxito cuando la página devuelta es la deseada. Si no es la deseada, todos los enlaces en ésta se agregan a la cola de prioridad con la relevancia estimada. La relevancia es

estimada usando un algoritmo recursivo de correspondencia de campo aplicado al contexto del enlace. El contexto del enlace es su texto ancla, o anchor text, y las dos líneas anteriores y las dos posteriores de la línea que contiene al texto.

Aunque cualquiera de las partes del proceso falle, el usuario recibe información útil. Si falla el primero, recibe la página de la institución del autor. Si falla el segundo, recibe la página de la institución del autor y la página personal del autor.

El principal problema que debe resolver WebFind es el problema de correspondencia de campo (field matching). Debe determinar si dos designadores sintácticamente diferentes son o no representaciones alternativas de una misma entidad, es decir si son o no semánticamente equivalentes. Por ejemplo, determinar si "UCSD" y "University of California, San Diego" son equivalentes. Este problema lo resuelve mediante un algoritmo.

El gran problema de este proyecto es que tiene una baja performance debido a que la búsqueda se realiza en tiempo real.

3.7. WebMate

WebMate [Chen, Sycara 1998] es un agente inteligente que ayuda a un usuario cuando navega y busca información en la web. Los motores de búsqueda no se adaptan a los intereses particulares de cada usuario, y WebMate intenta subsanar esto, manteniendo un perfil personalizado de los intereses del usuario.

Fue programado en Java. Los browsers, Netscape o Internet Explorer, necesitan ser configurados para usar WebMate como un servidor proxy http. El programa se puede bajar de la página de la Escuela de Ciencias de la Computación de la Carnegie Mellon University[13].

Las capacidades de WebMate a grandes rasgos son dos. La primera es aprender los intereses del usuario incrementalmente con una actualización continua y automáticamente proveerle de documentos que correspondan a su interés, como por

[13] www-2.cs.cmu.edu/~softagents/webmate.html

ejemplo un periódico personalizado. La segunda es ayudar al usuario a refinar la búsqueda para incrementar la recuperación de documentos relevantes.

La arquitectura de WebMate es una composición de un proxy stand-alone y un controlador applet. El proxy puede monitorear las acciones del usuario y aprender de ellas para proveer información para el aprendizaje y el refinamiento de búsquedas. A través del controlador applet, el usuario puede expresar sus intereses cuando navega y proveer un feedback de relevancia cuando busca. Adicionalmente, a través de éste, el usuario recibe ayuda inteligente de WebMate.

Con respecto al aprendizaje del perfil de usuario, WebMate lo realiza en forma automática, incremental y continua. Cuando el usuario marca un documento como de su interés, el sistema actualiza el perfil con esta información. De esta manera, se adapta a la evolución del usuario y a sus intereses recientes.

Este enfoque de aprender el perfil de usuario se utiliza para compilar un periódico personal. Esto se hace de dos formas. Una forma es controlar automáticamente una lista de URLs que el usuario indica y quiere que sean monitoreadas. Si el usuario no provee ninguna URL que quiere que sea la fuente de información, WebMate construye una consulta usando las palabras más utilizadas en su perfil actual y la manda a motores de búsqueda. Si se necesita el resultado inmediatamente, los resultados retornados por los motores de búsqueda son usados directamente como páginas recomendadas. Si no, el sistema va a buscar las páginas correspondientes a todas y cada una de las URLs en el resultado. Luego calcula la similitud del perfil y recomienda las páginas con una similitud mayor a un límite por orden de relevancia.

El agente WebMate, utiliza el contexto de las palabras de búsqueda en las páginas web relevantes para refinar la búsqueda. El fundamento de esto es que si el usuario le dice al sistema que una página es relevante a su búsqueda, el contexto de las palabras de búsqueda es más informativo que el contenido de la página. Es decir, dada una página relevante, el sistema primero busca por las palabras y por el contexto de estas palabras. El contexto de una palabra son las *n* palabras anteriores y las n palabras posteriores, con *n* a determinar. Se calculan las frecuencias de las palabras del contexto, y las mejor rankeadas, se usan luego para expandir las palabras utilizadas en la consulta.

3.8. Untangle

El proyecto Untangle[14] [Welty et al., 2000] aplica técnicas de Representación de Conocimiento y Razonamiento (KR&R) para el problema de encontrar información en la web.

Hay dos tecnologías claves que permiten trabajar a Untangle. La primera tecnología es una ontología para representar la información que está en forma electrónica, y una base del conocimiento implementada en la descripción de la lógica de Classic (CLASSification of Individual Concepts). Classic es un lenguaje de consulta propio. La segunda tecnología es una interfaz web para Classic, la cual permite a la base del conocimiento ser accedida interactivamente a través de cualquier browser web. La interfaz permite, para una consulta formulada en el lenguaje de consulta Classic, facilidades para la búsqueda más expresivas que cualquier herramienta de navegación actual.

El objetivo inicial del proyecto fue soportar inteligentemente la distribución de e-mail. Luego, con el crecimiento explosivo de la web los objetivos iniciales cambiaron, para proveer asistencia inteligente para la navegación en la web. El proyecto focalizó su primera fase en desarrollar una interfaz web para Classic. Esta interfaz visualizaba conceptos y descripciones individuales como páginas web dinámicas.

Untangle no presenta ningún descubrimiento nuevo. Es la aplicación de probar y conocer las verdaderas técnicas de representación para un dominio más visible: navegar la web. La contribución de este proyecto es demostrar que el uso de técnicas de KR&R aprovechables en la web produce beneficios prácticos.

La motivación original para trasladar esta investigación desde la distribución de mails a la web fue demostrar a la comunidad de Bibliotecas Digitales que las técnicas de KR pueden mejorar lo que se está haciendo en la recuperación de información. Pero esto va a ser difícil ya que las Bibliotecas Digitales y la web en general están fuertemente ligadas al área de la recuperación de información.

[14] untangle.cs.vassar.edu

3.9. Otros proyectos

En [Nagypàl, 2005] se propone una arquitectura para la recuperación de información que utiliza ontologías. El modelo de recuperación utilizado es el desarrollado para el proyecto Visual Contextualisation of Digital Content, VICODI[15], de la Unión Europea, que representa el contenido de un documento mediante un modelo temporal y otro similar al de espacio vectoriales. Los vectores de términos están formados por instancias de la ontología en lugar de las palabras del vocabulario de los documentos. Respecto a la expansión de consultas, inicialmente se aplican varias heurísticas sobre la consulta del usuario, para tratar con las imperfecciones que puedan presentarse en una ontología (falta de expresividad semántica, términos ambiguos, ontologías incompletas, etc.). Luego por cada una de las heurísticas aplicadas, se crean nuevas consultas independientes que son enviadas al motor de búsqueda. Por último, los resultados devueltos al usuario se forman combinando los resultados de cada una de las consultas independientes. Los metadatos se generan en forma automática, a partir de la ontología. Este proceso es realizado durante el indexado, con lo cual no se necesita insertar los metadatos en los documentos.

En [El-Beltagy et al., 2004] se propone una arquitectura para recuperar información de secciones individuales de documentos relativos a un dominio particular. Se utilizan metadatos para identificar dichas secciones, permitiendo a los usuarios realizar búsquedas estructuradas a partir de un conjunto predefinido de categorías que se mantienen en una ontología. El sistema está compuesto por un indexador, una ontología y un motor de bases de datos relacional. El indexador representa los documentos en XML y compara los campos XML y las categorías definidas en la ontología. Cuando hay coincidencia se crea un registro en una tabla de la base de datos. Al realizar las búsquedas se convierten las consultas del usuario al lenguaje de consulta SQL, y se resuelven en el motor relacional.

Swoogle [Ding et al., 2004] es un sistema de recuperación de información para documentos RDF y OWL que están en la web. Está diseñado para descubrir automáticamente tales documentos mediante un web crawler, indexar sus metadatos

y responder consultas acerca de los mismos. Además, el sistema tiene interfaces que le permiten interactuar con servicios web, personas y agentes de software.

Otra propuesta para realizar búsquedas en la web es usar un metabuscador. Los metabuscadores son servidores web que, dada una consulta del usuario, la envían a varios motores de búsqueda, reúnen las respuestas y las unifican. Ejemplos de metabuscadores pueden ser Metacrawler[16] y SavvySearch[17].

Las principales ventajas para un usuario al utilizar un metabuscador son utilizar una única interfaz común para realizar la misma consulta en distintas fuentes, y la habilidad del metabuscador de combinar los resultados mostrando una única respuesta. Los metabuscadores se diferencian unos de otros en cómo traducen la consulta del usuario al lenguaje de consulta específico de cada motor, y en cómo realizan el ranking en el resultado unificado. Este ranking contempla que las páginas retornadas por más de un motor son consideradas más relevantes.

Una desventaja de los metabuscadores es que cada uno de ellos tiene un conjunto de buscadores asociados. Por esto, la búsqueda no se envía a todos los motores de búsqueda. Entonces, puede suceder que el resultado no contenga necesariamente todas las páginas web que respondan a la consulta.

Los metabuscadores proveen los operadores AND, OR, ANDNOT y frase exacta; pero no preparan una estrategia de búsqueda adecuada, sino que dependen de la capacidad del usuario para escribirla. Es decir, no realizan el refinamiento semántico propuesto en esta tesis.

3.10. Comparación entre los distintos proyectos

En la página siguiente se presenta un cuadro comparativo (Tabla 3.1) de los proyectos analizados.

---

[15] www.vicodi.org/
[16] www.metacrawler.com
[17] www.SavvySearch.com

Proyecto	Descripción	Formalismo	Recursos utilizados	Enfoque de agentes	Refinamiento	Tipo de documento sobre el que actúa
CiteSeer	ACI orientado a buscar, indexar y recuperar papers	Modelo espacio vectorial Similitud de docs	Buscadores	Agentes	Expande la consulta a partir de las citas bibliográficas	Postscript, PDF
InfoSleuth	Recupera e integra información de fuentes heterogéneas		Ontologías	Red de agentes cooperantes	Utiliza las ontologías para mapear y para integrar, pero no para refinar la consulta	html
OntoBroker On2broker	Busca información e infiere respuestas en bases de datos cuya información es cargada a partir de páginas web	Frame Logic	Ontologías	La versión 2 utiliza agentes.	No realiza refinamiento semántico. Realiza inferencia sobre los datos	Bases de datos con información de págs web
OntoSeek	SRI basado en contenido.	LCG (grafos conceptuales léxicos). Compara grafos isomorfos	Ontologías	No	Desambiguación léxica y validación semántica.	Html y xml. Catálogos de productos y págs. amarillas on line.
Untangle	Recupera información de la web pero no con técnicas de IR, sino con técnicas de KR.	KR&R	Ontologías	No	No realiza refinamiento semántico. Utiliza ontologías para representar la estructura de los docs de la web	Html. Inicialmente para emails
WebFind	Descubre papers disponibles en la web por sus autores, en tiempo real	Similitud de docs	Bases de datos y fuentes de información externas: MelVyl, NetFind	No	Realiza la consulta detectando los autores de los documentos buscados y la amplía buscando papers de dichos autores relacionados con el tema.	Html. Páginas personales de los autores.
WebMate	SRI que asiste en la navegación de la web con perfiles personalizados de usuario.	Modelo espacio vectorial	Buscadores	Agentes Proxy: monitorea y aprende de las acciones del usuario Controlador applet: interactúa con el usuario	Expande la consulta utilizando términos obtenidos de un feedback de relevancia	Html

Tabla 3.1. Cuadro comparativo de proyectos relacionados

Como se ha mencionado en esta tesis, entre los recursos lingüísticos que pueden utilizarse como soporte para la recuperación de información están las ontologías. WordNet es considerado en muchos trabajos como una ontología, a pesar de no contar con axiomas. Varios de los proyectos analizados utilizan en forma general ontologías como recurso lingüístico. El uso de ontologías tiene numerosas ventajas, ya que permiten recuperación semánticamente correcta basándose en criterios específicos del dominio. Además, tanto su terminología como las relaciones entre términos se pueden actualizar.

Una característica en común que tienen varios de estos proyectos es el uso de agentes que utilizan ontologías como soporte para la búsqueda de información. El uso de agentes es muy importante, porque éstos conocen dónde buscar información y cómo obtenerla y proveen una interfaz expresiva e integrada para la web.

Una de las debilidades de los proyectos en vigencia relacionados con el tema, es que amplían la búsqueda en una sola dirección. Algunos lo hacen expandiendo los conceptos semánticamente. Muy pocos corrigen los conceptos ortográficamente sugiriéndole al usuario la forma ortográfica correcta del término. Ninguno le permite al usuario precisar su interés de búsqueda seleccionando un concepto jerárquicamente relacionado.

La propuesta de esta tesis para potenciar la recuperación de la información es ampliar la cantidad de documentos recuperados expandiendo el concepto semánticamente, previa verificación ortográfica del concepto a buscar. La verificación ortográfica tiene como objetivo evitar que los resultados sean erróneos o nulos, sugiriendo al usuario el término correcto. La expansión semántica incorpora a la búsqueda sinónimos a los fines de recuperar documentos que también sean relevantes aún cuando no respondan rigurosamente a las palabras utilizadas por el usuario, utilizando recursos lingüísticos del área del conocimiento.

Por otra parte, la mayoría de los sistemas analizados intentan automatizar completamente todas las tareas sin intervención del usuario. Por ejemplo, en los buscadores más populares suele suceder que ante una consulta simple se obtiene un gran número de documentos recuperados. Es de bien suponer que el usuario nunca podrá realizar una lectura del total con el objeto de clasificar cuáles pueden ser los documentos relevantes para su interés. Se propone mejorar la precisión a través de una

interacción mínima del usuario. Se requiere esta interacción para la desambiguación del concepto que permita presentarle al usuario la jerarquía de conceptos relacionada con la acepción de su interés, para que éste pueda incorporarlos a su consulta. El esfuerzo inicial que se pretende por parte del usuario es recompensado evitándole a posteriori la lectura y la clasificación manual de los documentos que no sean de su interés.

Capítulo 4: Refinamiento Semántico

4.1. Arquitectura propuesta

Como ya se ha presentado en el Capítulo 1, esta tesis se focaliza en aplicar los conceptos presentados en el Capítulo 2 teniendo en cuenta las debilidades discutidas en el Capítulo 3. Para ello se realiza el análisis y el desarrollo de un refinador semántico que utiliza recursos lingüísticos para construir una estrategia de búsqueda a partir de los conceptos ingresados por el usuario.

El *refinamiento semántico* que se propone consiste en guiar al usuario para *desambiguar* los conceptos ingresados por él, permitirle *seleccionar* conceptos jerárquicamente relacionados a fin de aumentar la precisión en los documentos a recuperar y *expandir* semánticamente los conceptos a fin de aumentar la cantidad de documentos a recuperar.

Para la *desambiguación* de conceptos, la solución propuesta es utilizar *recursos lingüísticos*, para que el usuario pueda decidir dentro de qué contexto se está buscando el concepto ingresado. Esta decisión la realiza en forma interactiva.

La *selección de conceptos jerárquicamente relacionados* consiste en mostrarle al usuario una jerarquía de conceptos vinculados con el concepto ya desambiguado, a fin de que el usuario se reubique, si es necesario, en una jerarquía conceptual para refocalizar su búsqueda y así aumentar la precisión en la recuperación. Esta etapa también es interactiva porque el usuario debe elegir los conceptos relacionados jerárquicamente provistos por el refinador a partir de los *recursos lingüísticos*.

La *expansión semántica* consiste en incorporar a la búsqueda términos que sean conceptualmente equivalentes: sinónimos y términos relacionados. Los sinónimos son grupos de palabras que representan un mismo concepto. Los términos relacionados son términos alternativos que, sin ser sinónimos ni estar relacionados jerárquicamente, pueden ser útiles para ampliar la cantidad de documentos a recuperar. Además, si el usuario desea obtener información en más de un idioma, entre estas expansiones se pueden incorporar la traducción de dichos términos. Esta expansión la realiza el refinador en forma automática mediante el uso de los *recursos lingüísticos*.

Finalmente, el resultado es una estrategia de búsqueda preparada en forma automática por el refinador. Una estrategia de búsqueda es una expresión lógica compuesta por distintos conceptos combinados con los conectores lógicos de conjunción, disyunción y negación.

La arquitectura propuesta para el refinamiento semántico se presenta en la Figura 4.1, donde los módulos sombreados indican que se necesita participación del usuario.

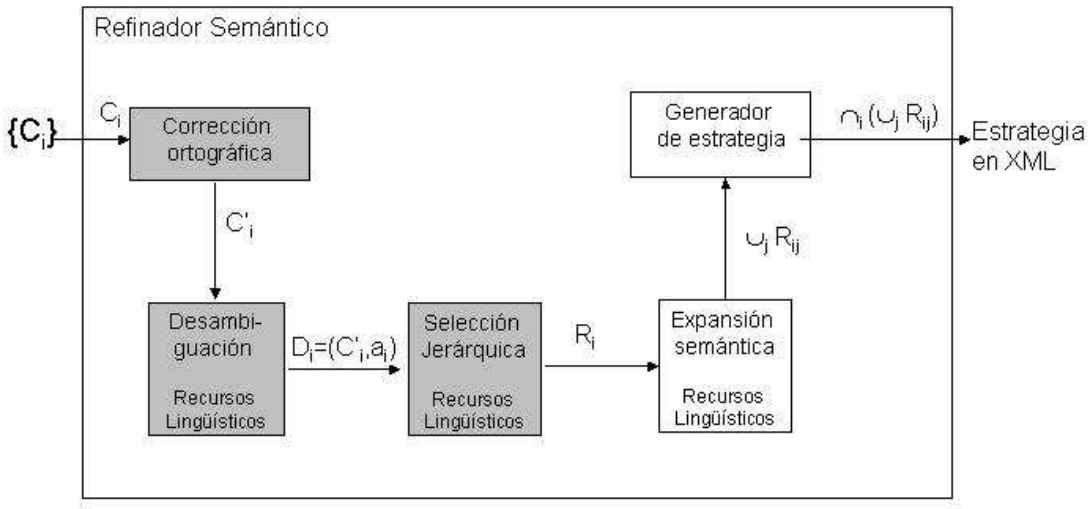


Figura 4.1: Arquitectura para el Refinamiento Semántico

Para realizar una consulta, el usuario ingresa un conjunto de conceptos $\{C_i\}$ con $1 \leq i \leq n$ y la salida del *Refinamiento Semántico* es una estrategia de búsqueda asociada a estos conceptos.

4.1.1. Corrección ortográfica:

Un primer paso en el armado de la estrategia es verificar que los términos estén correctamente escritos. Por cada concepto C_i que ingresa al módulo *Corrección ortográfica* se obtiene como salida un término corregido C'_i. Si C_i está bien escrito, C'_i

coincide con C_i. Si C_i estuviera incorrectamente escrito, entonces se lo reemplaza, previa aceptación del usuario, por C'_i.

4.1.2. Desambiguación:

La salida generada por el corrector ortográfico es luego procesada por el módulo *Desambiguación*. En este módulo, por cada concepto C'_i que ingresa se muestra al usuario las distintas acepciones asociadas al concepto, si las hubiera. El usuario selecciona la acepción que corresponde a su interés de búsqueda. Cada acepción de un concepto tiene una jerarquía conceptual asociada que es necesaria para los siguientes módulos. La salida de este módulo es el concepto D_i desambiguado de la forma (C'_i, a_i), donde C'_i es el concepto ingresado y a_i es la acepción elegida por el usuario.

Como se ha dicho en páginas anteriores, para realizar esta desambiguación se utilizan recursos lingüísticos tales como tesauros, diccionarios, diccionarios multilinguales y ontologías. Qué recurso o recursos utilizar, depende del área del conocimiento de la consulta y de los recursos disponibles para esa área. Un recurso disponible en línea muy utilizado para esta tarea es WordNet, y es el que se emplea en el prototipo.

La desambiguación que realiza el usuario permite continuar el proceso, en las etapas siguientes, con sólo aquellos términos que están vinculados conceptualmente con la acepción de interés del usuario.

4.1.3. Selección jerárquica:

El módulo *Selección jerárquica* muestra para cada concepto D_i los conceptos jerárquicamente relacionados con éste. Si existen conceptos jerárquicamente relacionados para algún D_i entonces, para aumentar la precisión de la búsqueda, se le permite al usuario moverse en la jerarquía conceptual de cada concepto D_i. Esto permite al usuario ubicar el concepto más cercano a su necesidad, permitiéndole *reemplazar* el concepto de partida D_i por algún otro concepto J_i que se encuentra jerárquicamente

relacionado en un nivel superior o inferior, o eventualmente en otra rama del árbol de jerarquía, y que represente más precisamente su interés de búsqueda. Si al usuario le interesa un conjunto de conceptos $J_{i,1}$, ..., $J_{i,s}$ de la jerarquía asociada al concepto D_i, la salida de este módulo es la *unión* de éstos. Es decir: $R_i = \bigcup_{j=1}^{s} J_{ij}$

Entonces, la entrada al módulo Selección jerárquica es D_i y la salida, que para simplificar la notación llamaremos R_i, puede ser:

- D_i si el usuario decidió no cambiar de nivel jerárquico;
- J_i si decidió reemplazar el concepto D_i, es decir el concepto C_i' con la acepción a_i, por otro concepto jerárquicamente relacionado; y
- $\bigcup_{j=1}^{s} J_{ij}$ si decidió reemplazar el concepto D_i, es decir el concepto C_i' con la acepción a_i, por un conjunto de conceptos jerárquicamente relacionados.

Generalmente, la tercera posibilidad indicada, se presenta cuando el usuario ingresa por un término general y le interesan dentro de éste varios hipónimos, es decir, varios términos específicos.

En este recorrido conceptual puede ocurrir que el usuario decida seleccionar un concepto específico, el cual sea ambiguo, es decir que pueda volver a tener más de una acepción. Por ejemplo, en el recurso WordNet ocurre que al buscar 'dog', y elegida por el usuario su acepción de animal, si selecciona el término específico 'sausage dog' dentro de la jerarquía, y dentro de este último el término específico 'barker', resulta que 'barker' tiene más de una acepción. 'Barker' además de ser un tipo de 'dog' es un término utilizado para 'Promoter' del área de marketing. Para no volver a requerir la participación del usuario, se automatiza esta desambiguación arrastrando la acepción original elegida por el usuario. En este ejemplo, se arrastra la acepción animal de 'dog'.

Para la selección jerárquica también se utilizan recursos lingüísticos, que pueden ser generales, como ser WordNet, o de un área específica del conocimiento, por ejemplo MeSH para el área salud. Los ejemplos mostrados están en inglés porque el recurso utilizado está en este idioma.

4.1.4. Expansión semántica:

La salida de la Selección jerárquica es procesada en el módulo *Expansión Semántica*, para encontrar sinónimos o términos relacionados para cada concepto R_i. Estas expansiones permiten aumentar la cantidad de documentos a recuperar. Entre estas expansiones también se pueden incorporar dichos términos en otros idiomas, si el usuario desea obtener información en más de un idioma. Los problemas que se presentan en la traducción de los conceptos se tratan en el Apéndice 2.

La salida de este módulo es un conjunto de *r* términos relacionados semánticamente $\{R_{i1}, ..., R_{ik} ... R_{ir}\}$ asociados a cada concepto R_i, con $1 \leq i \leq n$, donde n es la cantidad de conceptos que el usuario ingresa.

Es decir, la salida de la expansión semántica es: $\bigcup_{k=1}^{r} R_{ik}$

Entonces, para cada concepto C_i ingresado por el usuario al refinador, se obtiene el concepto C'_i corregido ortográficamente, luego el concepto D_i desambiguado, a continuación el concepto R_i jerárquicamente relacionado y finalmente, como resultado de la expansión semántica el conjunto $\bigcup_{k=1}^{r} R_{ik}$ de sinónimos y términos relacionados.

También aquí, se utilizan uno o más recursos lingüísticos para la incorporación de estos sinónimos.

4.1.5. Generación de estrategia:

Los conjuntos, formados por la unión de los R_{ik}, ingresan al *Generador de estrategia* cuya salida es la intersección de estas uniones, con $1 \leq i \leq n$, donde n es la cantidad de los conceptos ingresados. La salida del Generador de estrategia se representa en XML y contiene la estrategia de búsqueda asociada al interés del usuario.

Por lo tanto, este módulo escribe una estrategia que consiste en realizar en primer lugar el OR lógico de las expansiones semánticas de *cada* concepto; y luego el

AND lógico de estas expansiones.

Si el usuario desea hacer una búsqueda que *no* contenga un determinado concepto, se expande este concepto a descartar en la forma indicada en la arquitectura a fin de considerar otros sinónimos a descartar también. Luego se realiza el NOT del OR lógico obtenido para este concepto a negar y finalmente se lo agrega al AND lógico.

Es decir, para la búsqueda que involucra los conceptos

$$C_1 \text{ y } C_2 \text{ y .... y (no } C_h\text{) y ... y } C_n$$

planteada por el usuario, se obtiene la estrategia siguiente:

$$(R_{11} \text{ OR } R_{12} \text{ OR ... OR } R_{1r})$$

AND

.....

AND

$$(\text{NOT } (R_{h1} \text{ OR } R_{h2} \text{ OR ... OR } R_{hr}))$$

.....

AND

$$(R_{n1} \text{ OR } R_{n2} \text{ OR ... OR } R_{nr})$$

donde:

$(R_{11} \text{ OR } R_{12} \text{ OR ... OR } R_{1r})$ es la expansión del concepto C_1

...

$(\text{NOT } (R_{h1} \text{ OR } R_{h2} \text{ OR ... OR } R_{hr}))$ es la negación de la expansión del concepto C_h

...

$(R_{n1} \text{ OR } R_{n2} \text{ OR ... OR } R_{nr})$ es la expansión del concepto C_n

y el valor de *r* depende de cada concepto, pues todos los conceptos pueden no tener la misma cantidad de expansiones.

Esta estrategia se representa en XML resultando:

<estrategia>
    <concepto C_1>
        <ampliación 1> R_{11} </ampliación 1>
        .....
        <ampliación r> R_{1r} </ampliación r>
    </concepto C_1>
    <concepto C_2>
        <ampliación 1> R_{21} </ampliación 1>
        .....
        <ampliación r> R_{2r} </ampliación r>
    </concepto C_2>
    ......
    <no concepto C_h>
        <ampliación 1> R_{h1} </ampliación 1>
        .....
        <ampliación r> R_{hr} </ampliación r>
    </no concepto C_h>
    ......
    <concepto C_n>
        <ampliación 1> R_{n1} </ampliación 1>
        .....
        <ampliación r> R_{nr} </ampliación r>
    </concepto C_n>
</estrategia>

Figura 4.2: Estrategia genérica de búsqueda en XML

Este XML, resultante del refinador semántico es utilizado como entrada al siguiente módulo de la arquitectura general presentada en la Figura 1.1 del Capítulo 1. Este módulo es el *Adaptador de interfaz*, y se encarga de traducir la estrategia de búsqueda a la sintaxis de las distintas fuentes.

Qué recurso o recursos utilizar, depende del área del conocimiento de la consulta y de los recursos disponibles para esa área. Por ejemplo, para una consulta sobre temas médicos, se puede utilizar el tesauro MeSH[18].

4.2. Ventajas de automatizar la preparación de la estrategia de búsqueda

Las contingencias que se pueden encontrar en la preparación de una estrategia de búsqueda son: cómo reducir la cantidad si se recuperan demasiados documentos, y cómo aumentar la cantidad de documentos si no se recupera información suficiente.

En la recuperación de información tradicional, cuando un usuario común recupera demasiados documentos como resultado de una consulta, pudo haber cometido errores de estrategia o errores de entrada. Los errores de estrategia pueden provenir del uso de términos ambiguos o no específicos, de la falta de conceptos, del uso de disyunción (OR) cuando debería haber usado conjunción (AND) o del uso de truncamiento de términos demasiado corto. Los errores de entrada pueden deberse a un uso incorrecto de paréntesis.

En el caso de que el usuario común recupere pocos o ningún documento como resultado de una consulta, pudo haber cometido también errores de estrategia o errores de entrada. Los errores de estrategia en este caso pueden provenir del uso de demasiados conceptos, de no incluir sinónimos suficientes, de la utilización de términos demasiado específicos, del uso de operadores de proximidad sintáctica entre términos, del uso de conjunción (AND) cuando debe usarse disyunción (OR), o del uso incorrecto de la negación (NOT). Los errores de entrada en este caso pueden deberse a errores de tecleo, errores de deletreo (distintas formas de escribir una palabra, por ejemplo “center” en inglés americano y “centre” en inglés británico), o errores en paréntesis.

El refinador semántico resuelve la mayoría de estos problemas: la desambiguación de términos ambiguos o no específicos, el correcto uso de la disyunción y de la conjunción, el uso correcto de paréntesis, la inclusión de sinónimos y palabras con distintas formas de escritura, la utilización de términos específicos, el uso

[18] www.nlm.nih.gov/mesh/meshhome.html

correcto de la negación y los errores de tecleo. De esta forma, se mejora la recuperación de información.

4.3. Ejemplos

En esta sección se describe la utilización de la arquitectura planteada en casos de uso. En primer lugar se resuelve con la arquitectura propuesta el ejemplo motivador planteado en la Sección 1.2. del Capítulo 1. En segundo lugar, se resuelve una variante de este ejemplo, en la cual el usuario refocaliza su interés de búsqueda a partir de la estructura jerárquica de conceptos. Finalmente, se presenta un tercer ejemplo en el cual la búsqueda involucra varios conceptos.

Ejemplo 1

En este ejemplo un usuario médico desea obtener información sobre “cáncer de pulmón”. Debido a que la mayor parte de información científica del área salud está en idioma inglés, y a que la mayoría de los recursos lingüísticos de esta área también lo están, el usuario decide realizar la consulta en inglés, y decide ingresar el concepto más general “cancer”.

El refinador semántico toma esta palabra y verifica que está correctamente escrita desde el punto de vista ortográfico. Si el usuario hubiera ingresado “canser”, el corrector le sugiere la palabra “cancer” ortográficamente correcta.

La palabra “cancer” ingresa al módulo Desambiguación el cual a través de un recurso lingüístico le muestra las distintas acepciones de esa palabra. Si se utiliza WordNet como recurso lingüístico, se observa que el sistema provee cinco acepciones distintas de esta palabra (Figura 4.3).

En este ejemplo queda evidente que la semántica del concepto depende del contexto en el cual es usado, o dicho de otra forma, del dominio de la aplicación.

El usuario decide que la acepción de interés es la primera. El módulo selección de jerarquía expande entonces este concepto con sus hipónimos. (Figura 4.4).

The **noun** "cancer" has 5 senses in WordNet.

1. **cancer**, malignant neoplastic disease -- (any malignant growth or tumor caused by abnormal and uncontrolled cell division; it may spread to other parts of the body through the lymphatic system or the blood stream)
2. Cancer, Crab -- ((astrology) a person who is born while the sun is in Cancer)
3. Cancer -- (a small zodiacal constellation in the northern hemisphere; between Leo and Gemini)
4. Cancer, Cancer the Crab, Crab -- (the fourth sign of the zodiac; the sun is in this sign from about June 21 to July 22)
5. Cancer, genus Cancer -- (type genus of the family Cancridae)

Figura 4.3: Respuesta de WordNet para el término "cancer"

Results for "Hyponyms (...is a kind of this), full" search of noun "cancer"

Sense 1
cancer, malignant neoplastic disease --
(any malignant growth or tumor caused by abnormal and uncontrolled cell division; it may s
pread to other parts of the body through the lymphatic system or the blood stream)
=> lymphoma --
(a neoplasm of lymph tissue that is usually malignant; one of the four major types of cancer)
=> carcinoma --
(any malignant tumor derived from epithelial tissue; one of the four major types of cancer)
=> liver cancer, cancer of the liver --
(malignant neoplastic disease of the liver usually occurring as a metastasis from another canc
er; symptoms include loss of appetite and weakness and bloating and jaundice and upper abd
ominal discomfort)
=> adenocarcinoma, glandular cancer, glandular carcinoma --
(malignant tumor originating in glandular epithelium)
=> prostate cancer, prostatic adenocarcinoma -- (cancer of the prostate gland)
=> breast cancer --

(cancer of the breast; one of the most common malignancies in women in the US)
=> carcinoma in situ, preinvasive cancer --
(a cluster of malignant cells that has not yet invaded the deeper epithelial tissue or spread to other parts of the body)
=> colon cancer -- (a malignant tumor of the colon; early symptom is bloody stools)
=> **lung cancer** -- (carcinoma of the lungs; one of the commonest forms of cancer)
=> pancreatic cancer -- (cancer of the pancreas)
=> leukemia, leukaemia, leucaemia, cancer of the blood --
(malignant neoplasm of blood-forming tissues; characterized by abnormal proliferation of leukocytes; one of the four major types of cancer)
=> acute leukemia -- (rapidly progressing leukemia)
=> acute lymphocytic leukemia, acute lymphoblastic leukemia --
(acute leukemia characterized by proliferation of immature lymphoblast-like cells in bone marrow, lymph nodes, spleen, and blood; most common in children)
…..

Figura 4.4: Respuesta de WordNet para la selección de hipónimos de "cancer"

El usuario se mueve en la jerarquía y se queda con la frase "lung cancer", la cual ingresa al módulo Expansión semántica. Este módulo expande por sinónimos y términos relacionados sin intervención del usuario.

Si para esta expansión se utilizara el recurso WordNet, éste provee el siguiente conjunto de sinónimos (Figura 4.5):

Results for "Synonyms, ordered by estimated frequency" search of noun "lung cancer"

Sense 1
lung cancer -- (**carcinoma of the lungs**; one of the commonest forms of cancer)
=> carcinoma --
(any malignant tumor derived from epithelial tissue; one of the four major types of cancer)

Figura 4.5: Respuesta de WordNet para la expansión por sinónimos de "lung cancer"

Por lo tanto, el módulo Expansión Semántica incorpora automáticamente el término: “carcinoma of the lungs”.

Si en el módulo Expansión semántica, además de utilizar el recurso WordNet se utilizan otros recursos tales como un diccionario multilingual y un tesauro, por ejemplo MeSH (Medical Subject Subheadings), se incorporan otros conceptos tales como: “cáncer de pulmón”, obtenido a partir del diccionario multilingual, y “lung neoplasms”, obtenido del tesauro MeSH.

El módulo Generador de estrategia, toma el término seleccionado en la jerarquía por el usuario “lung cancer”, y sus sinónimos y, en forma automática, construye la estrategia de búsqueda.

Entonces, para este ejemplo:

$C_1 = C_1' =$ cancer

$D_1 =$ (cancer, “malignant neoplastic disease”)

$J_1 =$ lung cancer

$R_1 = J_1 =$ (lung cancer, “malignant neoplastic disease”)

$R_{11} =$ lung cancer

$R_{12} =$ carcinoma of the lungs

$R_{13} =$ cáncer de pulmón

$R_{14} =$ lung neoplasms

Por lo tanto:

$\bigcup_{k=1}^{r} R_{1k}$ = { lung cancer, carcinoma of the lungs, cáncer de pulmón, lung neoplasms }

y la estrategia de búsqueda obtenida es:

lung cancer OR carcinoma of the lungs

OR cáncer de pulmón OR lung neoplasms

Ejemplo 2

Supongamos ahora que un usuario desea obtener información sobre un “tipo particular de cáncer de pulmón”. Debido a que la mayor parte de información científica del área salud está en idioma inglés, y a que la mayoría de los recursos lingüísticos de esta área también lo están, el usuario decide realizar la consulta en inglés. Como desconoce el término exacto en inglés para este subtipo de cáncer de pulmón, decide entonces ingresar el concepto más general “lung cancer”.

El módulo Corrección ortográfica del refinador semántico toma esta frase y verifica que está correctamente escrita desde el punto de vista ortográfico.

La frase “lung cancer” ingresa al módulo Desambiguación el cual a través de un recurso lingüístico le muestra las distintas acepciones de esa palabra. Si se utiliza WordNet como recurso lingüístico, se observa que este recurso provee para esta frase una única acepción (Figura 4.6).

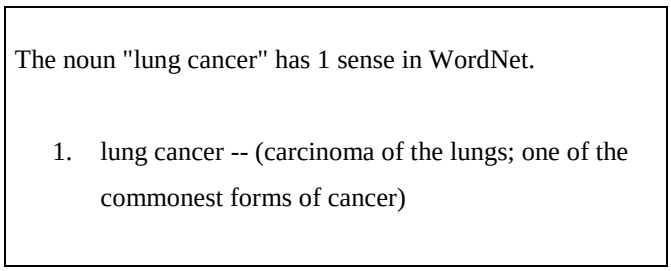
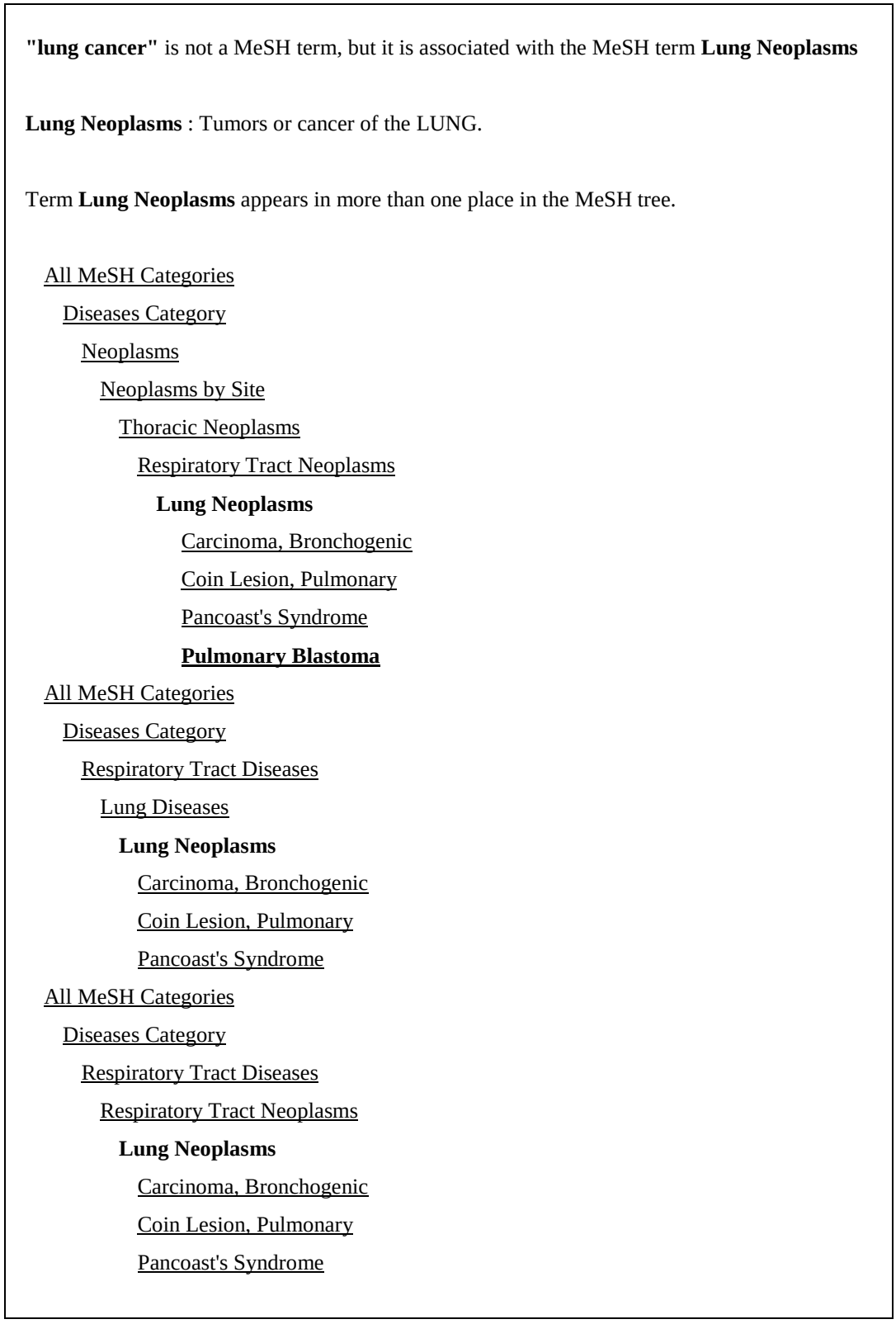


Figura 4.6: Respuesta de WordNet para el término “lung cancer”

En este caso, no es necesario desambiguar el término porque tiene una única acepción. El módulo Selección de jerarquía expande entonces este concepto mostrando los conceptos jerárquicamente relacionados. Si para esto se utiliza como recurso el tesauro MeSH, al ser éste un término prohibido, el tesauro refiere en forma automática a su término permitido: “lung neoplasms”, mostrando además los conceptos relacionados jerárquicamente con éste último. Como puede observarse en la Figura 4.7, un término MeSH puede aparecer en varias jerarquías conceptuales, y el usuario puede moverse por estas jerarquías para ubicar el concepto más cercano a su necesidad. Al ver las jerarquías mostradas, el usuario reconoce que su término de interés es “pulmonary blastoma”. Entonces, la posibilidad de moverse por estas jerarquías, subiendo o bajando

de nivel conceptual, permite al usuario precisar mejor su búsqueda.



"lung cancer" is not a MeSH term, but it is associated with the MeSH term **Lung Neoplasms**

Lung Neoplasms : Tumors or cancer of the LUNG.

Term **Lung Neoplasms** appears in more than one place in the MeSH tree.

All MeSH Categories
Diseases Category
Neoplasms
Neoplasms by Site
Thoracic Neoplasms
Respiratory Tract Neoplasms
Lung Neoplasms
Carcinoma, Bronchogenic
Coin Lesion, Pulmonary
Pancoast's Syndrome
Pulmonary Blastoma

All MeSH Categories
Diseases Category
Respiratory Tract Diseases
Lung Diseases
Lung Neoplasms
Carcinoma, Bronchogenic
Coin Lesion, Pulmonary
Pancoast's Syndrome

All MeSH Categories
Diseases Category
Respiratory Tract Diseases
Respiratory Tract Neoplasms
Lung Neoplasms
Carcinoma, Bronchogenic
Coin Lesion, Pulmonary
Pancoast's Syndrome

Figura 4.7: Una vista del tesauro MeSH de Medline

Esta frase, elegida por el usuario, reemplaza a la frase de partida “lung cancer” y es ingresada al módulo Expansión semántica. Este módulo incorpora automáticamente las frases “pulmonary blastomas”, utilizando como recurso un diccionario, y “blastoma pulmonar”, utilizando como recurso un diccionario multilingual.

Finalmente, el módulo Generador de estrategia, construye la estrategia de búsqueda.

Para este ejemplo:

$C_1 = C_1' =$ lung cancer

$D_1 =$ (lung cancer, “tumors or cancer of the lung”)

$J_1 =$ pulmonary blastoma

$R_1 = J_1$

$R_{11} =$ pulmonary blastoma

$R_{12} =$ pulmonary blastomas

$R_{13} =$ blastoma pulmonar

Por lo tanto:

$$\bigcup_{k=1}^{r} R_{1k} = \{ \text{pulmonary blastoma, pulmonary blastomas, blastoma pulmonar} \}$$

y la estrategia final de búsqueda es:

pulmonary blastoma OR pulmonary blastomas OR blastoma pulmonar

Ejemplo 3

Los ejemplos anteriores son sencillos pues el usuario ha planteado su necesidad de información a partir de *un solo* concepto. Pero, generalmente una búsqueda involucra varios conceptos. En estos casos, el refinador semántico trata cada uno de estos conceptos en forma independiente, como se muestra en los ejemplos anteriores, y sus expansiones se combinan en el módulo Generación de estrategia. Como resultado, la estrategia de búsqueda asociada consta de la disyunción de cada una de las expansiones y luego la conjunción de los conjuntos resultantes de las expansiones.

Por ejemplo, si se desea saber la "relación de la aspirina en el tratamiento del cáncer de pulmón".

Los conceptos que ingresa el usuario son: *cáncer de pulmón* - *aspirina* - *tratamiento*. Por cada uno de estos conceptos, el refinador realiza un procedimiento similar al mostrado en los ejemplos anteriores.

La estrategia de búsqueda final provista por el Generador de estrategia es:

(lung neoplasms OR lung cancer
OR cáncer de pulmón OR carcinoma of the lungs)
AND
(aspirina OR aspirin OR ácido acetil salicílico)
AND
(tratamiento OR treatment)

Como se mencionó al comienzo de este capítulo, estas estrategias son representadas en el formato estándar de intercambio de datos XML. A modo de ejemplo, esta última estrategia de búsqueda, representada en XML, es la mostrada en la Figura 4.8.

```
<estrategia>
    <concepto 1>
        <ampliación 1>cáncer de pulmón</ampliación 1>
        <ampliación 2>lung cancer</ampliación 2>
        <ampliación 3>lung neoplasms</ampliación 3>
        <ampliación 4>carcinoma of the lungs </ampliación 4>
    </concepto 1>
    <concepto 2>
        <ampliación 1>aspirina</ampliación 1>
        <ampliación 2>aspirin</ampliación 2>
        <ampliación 3>ácido acetil salicílico</ampliación 3>
    </concepto 2>
    <concepto 3>
        <ampliación 1>tratamiento</ampliación 1>
        <ampliación 2>treatment</ampliación 2>
    </concepto 3>
</estrategia>
```

Figura 4.8: Estrategia de búsqueda en XML para el ejemplo 3

4.4. Prototipo

Para el desarrollo del prototipo se utilizaron estándares y recomendaciones del grupo W3C así como también lenguajes y recursos libres disponibles en la web. En primer lugar, se buscó en la web qué recursos computacionales y/o de información estaban disponibles para utilizar en el prototipo. Para la corrección ortográfica se utiliza el método Spelling Suggestion del Web Service de Google[19]. Para la selección de jerarquía y la expansión semántica se utiliza el recurso lingüístico WordNet, por ser uno de los más utilizados en trabajos similares. Se utiliza la versión 1.6 en formato RDF, por limitaciones de espacio de almacenamiento en el servidor disponible.

[19] www.google.com/apis

Estos servicios fueron ensamblados y provistos de una interfase web sencilla. Se adoptó PHP[20] como lenguaje para implementar el prototipo, dado que es un recurso libre, es soportado por los servidores web utilizados en este desarrollo, que están bajo plataforma GNU/Linux, y brinda la posibilidad de trabajar con modelos simples de objetos.

Para la transformación de esquemas se optó por XSLT[21] [W3C, 1999] [Kay, 2001]. La elección de XSLT obedece a criterios técnicos específicos. Dado que el recurso lingüístico con el que se interactúa trabaja con datos en formato XML, y en vistas que estos datos deben ser procesados, XSLT resulta ser la mejor opción para convertir a su presentación en Html.

Se analizaron distintos buscadores encontrando que Google[22] tenía en el momento de la experimentación la limitación de 10 palabras por consulta, y una estrategia compleja puede llegar a tener muchas palabras más. Se analizó el buscador Yahoo![23] y se observó que no tenía estas limitaciones. Por eso se utilizó este último buscador en las experiencias.

El prototipo se ha desarrollado tomando como base el modelo de arquitectura propuesto, y fue implementado por el Ing. Jorge Saer dentro del Proyecto de Investigación bajo mi dirección: “Recuperación de Información Basada en Semántica” [Deco, 2004-2006].

La entrada al prototipo es un conjunto de conceptos ingresados por el usuario y la salida del prototipo es una estrategia de búsqueda.

La Figura 4.9 muestra la pantalla inicial del mismo. En la versión actual se requiere que el concepto a buscar esté en inglés. Esto es una limitación por el uso de WordNet como recurso libre de la web.

[20] www.php.net
[21] www.w3.org/TR/xslt
[22] www.google.com
[23] www.yahoo.com

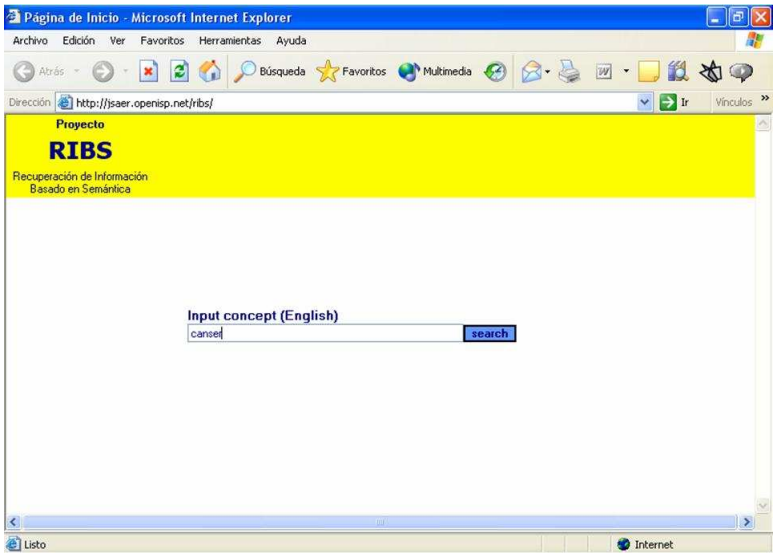


Figura 4.9. Pantalla inicial del prototipo de refinador semántico

A partir de la pantalla de la Figura 4.9, los pasos a seguir son:

- El usuario ingresa, en inglés, el concepto a buscar.
- Con el botón de búsqueda “search”, envía al sistema la palabra ingresada.
- El sistema presenta la pantalla del corrector ortográfico.
    - Si el término ingresado es correcto, aparece un cartel invitando a continuar el proceso sobre esa palabra. El usuario debe seleccionar ese término, que aparecerá como hipervínculo, para continuar.
    - Si el término ingresado es considerado por el corrector ortográfico como inexistente o mal escrito, se le ofrece al usuario la opción de continuar con un término cuya grafía es aproximada a la ingresada y que sí aparece como correcto según el corrector.
- El desambiguador muestra las diferentes acepciones del concepto, si es que éste tiene más de una acepción.
- El usuario debe seleccionar una acepción del término en cuestión.
- El sistema despliega una pantalla en donde, en forma de título, se muestra el término, seguido de su hiperónimo y por debajo se muestra una lista de sus hipónimos.

- El usuario selecciona los términos de su interés de esta jerarquía conceptual.
- El sistema amplía automáticamente cada uno de éstos agregándole sus sinónimos.
- El sistema prepara la estrategia de búsqueda asociada.

Si la búsqueda involucra varios conceptos, este proceso se realiza por cada uno de estos conceptos. En el Apéndice 3 se describe el prototipo con mayor detalle.

Capítulo 5: Experimentación con la arquitectura propuesta utilizando recursos lingüísticos generales y específicos

El refinamiento semántico propuesto tiene por objetivo formular una estrategia de búsqueda a partir de conceptos ingresados por el usuario. Este refinamiento resuelve muchos de los problemas que se presentan en la formulación de una estrategia de búsqueda, como ser el correcto uso de la disyunción y de la conjunción, el uso correcto de paréntesis, la inclusión de sinónimos y palabras con distintas formas de escritura, la utilización de términos específicos, el uso correcto de la negación y los errores de tecleo. Mediante la formulación de una estrategia de búsqueda correcta, es posible aumentar la cantidad de documentos recuperados y la precisión de los resultados.

La cantidad de documentos recuperados aumenta si se amplía en forma automática el criterio de búsqueda ingresado por el usuario, mediante el agregado de sinónimos y palabras relacionadas. La precisión de los resultados se logra presentándole una estructura jerárquica de conceptos que le permite hacer un recorrido conceptual de su consulta. Es decir, moverse por estas jerarquías, subiendo o bajando de nivel conceptual, y seleccionando un término más preciso a su necesidad de información.

En esta tesis se propone la utilización de distintos recursos lingüísticos, tales como tesauros u ontologías para este fin. La Sección 5.1. muestra la experimentación realizada a partir de la arquitectura propuesta para búsquedas de interés general, utilizando un recurso lingüístico de cobertura general. En la Sección 5.2. se describe la experimentación realizada para consultas en un dominio específico del conocimiento, para lo cual se utiliza un recurso lingüístico especializado en ese dominio y se discuten trabajos relacionados en dominios específicos. Finalmente, en la Sección 5.3. se presentan conclusiones sobre ambas experimentaciones.

5.1. Experimentación en un dominio general

Un recurso lingüístico de cobertura general disponible en línea y con el cual se

realizaron las experiencias que se describen en esta sección, es WordNet[24].

Este recurso incluye sinónimos, variantes de escritura de nombres propios, ampliación de siglas, variaciones de deletreo, y para ciertos términos su escritura en otros idiomas. Por este motivo se lo consideró adecuado para la experimentación, dado que el agregado de estos términos permite aumentar la cantidad de documentos recuperados.

Además, WordNet tiene una jerarquía conceptual, y muestra para cada término sus términos específicos o hipónimos, y su término más amplio o hiperónimo. Por este motivo también se lo consideró adecuado para la experimentación, porque el usuario puede recorrer esta jerarquía, subiendo o bajando de nivel conceptual, y seleccionar un término más preciso a su necesidad de información. Con lo cual se mejora la precisión de los resultados.

Para probar el refinamiento semántico, se realizaron 24 consultas. Para cada consulta se solicitó al usuario que describiera su interés de búsqueda en sus propias palabras, y que luego realizara la consulta de dos formas: primero en el buscador Yahoo! y luego con el refinamiento semántico. Se registró la estrategia planteada por el usuario directamente a Yahoo! y se registró la estrategia generada por el refinamiento semántico, que luego se ejecutó en Yahoo!. Además, en cada prueba se registró la cantidad de documentos resultantes y la cantidad de documentos que respondían al interés del usuario en los primeros 50 documentos, a fin de medir luego la precisión en los primeros 50 documentos. Además se registró el tipo de usuario que realizaba la consulta. Se consideró de nivel Inexperto a aquel usuario que no estaba habituado al uso de un buscador. El nivel Medio corresponde a los usuarios que realizan consultas a través de buscadores con frecuencia. Un usuario de nivel Experto es aquel que utiliza las opciones de Búsqueda Avanzada en los buscadores, o aquel que utiliza, y en forma adecuada, operadores lógicos en su consulta.

En la Tabla 5.1 se presentan las consultas realizadas. A continuación se presentan observaciones sobre la cantidad de documentos recuperados y sobre la cantidad de documentos relevantes en los primeros 50 documentos.

[24] wordnet.princeton.edu/

		Búsqueda sin refinamiento en Yahoo!			Búsqueda con refinamiento en Yahoo!				
Nro. consulta	Descripción	Expresión búsqueda	Cantidad docs. resultantes	Cant. docs. relevantes (primeros 50 docs.)	Términos para el refinamiento semántico	Estrategia de búsqueda	Cantidad docs. resultantes	Cant. docs. relevantes (primeros 50 docs.)	Nivel de usuario
1	Países que componen la Comunidad Económica Europea	countries of the eec	225.000	10	- EEC - countries	(“European Union” OR EU OR “European Community” OR EC OR “European Economic Community” OR EEC OR “Common Market Europe”) AND (country OR state OR land)	18.600.000	28	Inexperto
2	Pinturas de Salvador Dalí	Dali’s pictures	18.000	23	- Dali - pictures	(dali OR “salvador dali”) AND (picture OR painting)	459.000	36	Inexperto
3	Biografía de Mendel	Mendel	590.000	29	- Mendel	Mendel	590.000	29	Medio
4	Ganadores del premio Nobel de Medicina	nobel + medicine + winners	88.900	31	- nobel - medicine - winners	(Nobel OR “Alfred Nobel” OR “Alfred Bernhard Nobel”) AND (medicine OR “medical specialty”) AND (achiever OR winner OR success OR succeeder)	258.000	11	Medio
5	Ganadores del premio Nobel de Medicina	nobel + medicine + winners	88.900	31	- nobel prize - medicine - winners	“nobel prize” AND (medicine OR “medical specialty”) AND (achiever OR winner OR success OR succeeder)	148.000	21	Medio
6	Libros escritos por García Márquez	gabriel garcia marquez books	206.000	35	- book - Gabriel García Márquez	“gabriel garcia marquez” AND book	165.000	41	Inexperto
7	manual del vehículo Peugeot Partner	Peugeot Partner manual	29.800	9	- Peugeot Partner - manual	“Peugeot Partner” AND (manual OR handbook)	1.180	14	Inexperto

Tabla 5.1: Consultas realizadas con y sin refinamiento

		Búsqueda sin refinamiento en Yahoo!			Búsqueda con refinamiento en Yahoo!				
Nro. consulta	Descripción	Expresión búsqueda	Cantidad docs. resultantes	Cant. docs. relevantes (primeros 50 docs.)	Términos para el refinamiento semántico	Estrategia de búsqueda	Cantidad docs. resultantes	Cant. docs. relevantes (primeros 50 docs.)	Nivel de usuario
8	Qué es Escherichia Colli?	Escherichia Colli	680	9	- Escherichia Colli	"Escherichia Coli"	1.090.000	46	Medio
9	ejemplos de data mining	"data mining" example	330.000	21	- data mining - example	"data mining" AND (exercise OR example)	359.000	21	Experto
10	Tratamientos de linfoma de Hodgkin	hodgkin's lymphoma treatments	141.000	10	- lymphoma - treatment - hodgkin	"hodgkin's disease" AND (treatment OR "medical care" OR "medical aid") AND (Hodgkin OR "Thomas Hodgkin")	208.000	35	Medio
11	Año en que se desarrolló la Guerra Civil Española	"Spanish Civil War"	230.000	30	- spanish - civil - war	spanish AND (civil OR civic) AND (war OR warfare)	2.230.000	3	Experto
12	Año en que se desarrolló la Guerra Civil Española	"Spanish Civil War"	230.000	30	- war	"Spanish civil war"	230.000	30	Experto
13	Sistemas de comunicaciones GSM	GSM communication	1.280.000	39	- GSM - communication	GSM AND communication	1.280.000	39	Medio
14	Cuáles son las especies que se encuentran en extinción	Species in extinction	1.040.000	31	- species - extinction	species AND (extinction OR defunctness)	1.050.000	29	Inexperto
15	Tipos y lugares de carreras de karting	Karting Race	431.000	21	- Karting Race	"Karting Race"	3.660	26	Inexperto
16	Qué es el complejo de Edipo	Oedipus Complex	87.900	22	- Oedipus Complex	"Oedipus Complex" OR "Oedipal Complex"	53.100	33	Medio

Tabla 5.1: Consultas realizadas con y sin refinamiento (cont.)

		Búsqueda sin refinamiento en Yahoo!			Búsqueda con refinamiento en Yahoo!				
Nro. consulta	Descripción	Expresión búsqueda	Cantidad docs. resultantes	Cant. docs. relevantes (primeros 50 docs.)	Términos para el refinamiento semántico	Estrategia de búsqueda	Cantidad docs. resultantes	Cant. docs. relevantes (primeros 50 docs.)	Nivel de usuario
17	Tratamientos de bulimia	bulimia treatment	403.000	47	- treatment - bulimia	treatment AND (bulimia OR "binge-eating syndrome")	408.000	47	Medio
18	Distribuidores de software en Estados Unidos	software distributor usa	941.000	23	- software - distributor - usa	(software OR "software system" OR "software package" OR package) AND (distributor OR distributer) AND (usa OR "u.s.a." OR us OR "u.s." OR "United States of America" OR "United states")	2.250.000	21	Medio
19	Distribuidores de software en Estados Unidos	software distributor usa	941.000	23	- software - distributor - usa	(software OR "software system" OR "software package" OR package) AND (distributor OR distributer) AND ("United States of America" OR "United states")	1.520.000	25	Medio
20	Comidas que no contienen azúcar	food without sugar	2.680.000	13	- food	"diabetic diet"	89.800	19	Inexperto
21	Modelos de Yamaha Virago	Yamaha virago model	34.400	14	- yamaha virago - model	"yamaha virago" AND model	7.800	48	Inexperto
22	Descubrimiento de la penicilina	penicillin discovery	76.200	7	- penicillin - discovery	penicillin AND (discovery OR find OR uncovering)	78.900	7	Medio
23	Costo de licencia de SQL server	Cost license "SQL server"	88.700	26	- cost - license - sql server	(cost OR "monetary value" OR price) AND (license OR licence OR permit) AND "SQL server"	282.000	17	Medio
24	Terapia de la depresión	therapy of the depression	2.630.000	23	- therapy - depression	(therapy OR psychotherapy OR psychoanalisis) AND (depression OR melancholia OR melancholy OR dejection)	2.880.000	37	Inexperto

Tabla 5.1: Consultas realizadas con y sin refinamiento (cont.)

5.1.1. Cantidad de documentos recuperados

- *Consultas con nombres propios y siglas*

Con el refinamiento semántico, en las consultas 1, 2, 4, 5, 18 y 19 se observa que aumenta notablemente la cantidad de documentos recuperados. Esto se debe a que en el caso de utilizar nombres propios o siglas, si éstos existen en WordNet, se amplía la estrategia de búsqueda con sus variantes de escritura.

En consultas como la 3, donde *Mendel* también es un nombre propio pero que no está en WordNet, y la 13, donde *GSM* es una sigla que no está en WordNet, no cambian los valores obtenidos con respecto a la consulta sin refinamiento semántico en el buscador Yahoo!.

- *Consultas con frases*

En la consulta 6, el nombre propio *Gabriel García Márquez* no se encuentra en WordNet, pero el refinamiento semántico agrega a esta frase las comillas: "*Gabriel García Márquez*". El agregado de las comillas hace que el buscador lo considere como frase y no como tres palabras independientes, como lo toma en la consulta sin refinar del usuario. Esto explica la reducción de la cantidad de documentos recuperados. Algo similar ocurre en la consulta 7 con la frase *Peugeot Partner*, que corresponde a un nombre propio que no se encuentra en WordNet, y en la consulta 15 donde la frase *Karting Race* tampoco se encuentra en WordNet. Algo similar ocurre en la consulta 21.

- *Consultas con errores ortográficos*

En consultas como la 8, la cantidad de documentos resultantes sin el refinamiento semántico es notablemente menor que con el refinamiento semántico. Esto se debe a que la palabra ingresada, *colli*, estaba mal escrita. En este caso, el refinamiento ofrece la posibilidad de buscar la palabra *coli*, la cual es el término ortográficamente correcto.

- *Consultas con términos con pocos o ningún sinónimo en WordNet*

Se puede observar en la consulta 9 que la cantidad de documentos recuperados con refinamiento y sin refinamiento es prácticamente la misma. Esto se debe a que los términos buscados *data mining* y *example*, están en WordNet pero el primero no tiene sinónimos y el segundo aporta un solo sinónimo. Algo similar ocurre en las consultas 14, 17 y 22.

- *Consultas con términos más específicos*

En la consulta 10, la cantidad de documentos recuperados con refinamiento también aumenta. Esto se debe al agregado de sinónimos a los términos *treatment* y *Hodgkin*. Sin embargo, esta cantidad de documentos recuperados no ha sido mucho mayor a los obtenidos sin refinamiento semántico, dado que el usuario ingresa por el término *lymphoma*, se mueve por la jerarquía conceptual asociada y decide reemplazar este término de partida por *"hodgkin's disease"*, que es un tipo específico de linfoma y que responde mejor a su interés. En la consulta 20, el usuario al ingresar al refinador, se movió por la jerarquía del término *food* y decidió quedarse con un término más específico "*diabetic diet"*. Esto disminuyó la cantidad de documentos recuperados.

Las consultas 11 y 12 corresponden a un mismo interés de búsqueda y están resueltas con dos formas distintas de realizar el refinamiento semántico. En la consulta 11, el usuario ingresa tres términos: *spanish*, *civil* y *war*, y el refinamiento amplía cada uno de éstos con los respectivos sinónimos. En la consulta 12, el usuario decide ingresar por el término *war* y recorriendo la jerarquía conceptual baja de nivel a "*civil war"* y dentro de éste, baja nuevamente de nivel para optar por un tipo particular de guerra civil: *"spanish civil war"*.

En la consulta 11, la cantidad de documentos recuperados con refinamiento semántico es mucho mayor a la obtenida sin refinamiento semántico. En cambio, en la consulta 12 la cantidad de documentos resultantes es idéntica con y sin refinamiento semántico. Una posible explicación de esto es que en la consulta 11, el usuario ingresó como conceptos para el refinamiento los adjetivos *spanish* y *civil*, que en realidad no son conceptos sino adjetivos calificativos de *war*. Lo correcto sería en casos como éste, realizar una estrategia como la de la consulta 12, donde se ingresa por el sustantivo

principal y se eligen, recorriendo la jerarquía conceptual, como términos específicos los sustantivos adjetivados.

Un caso similar se presenta en las consultas 4 y 5, donde es distinto buscar a *Nobel* como persona, como en la consulta 4 donde WordNet ofrece *"Alfred Nobel"* y *"Alfred Bernhard Nobel"* como sinónimos, o buscar *"nobel prize"*, como en la consulta 5, donde *"nobel prize"* es un tipo de *prize*.

Además, en la consulta 12 no hay diferencia entre la estrategia resultante del refinamiento semántico y la estrategia sin refinamiento escrita por el usuario. Esto se debe a que el usuario es experto e ingresa de entrada las tres palabras como una sola frase.

En la consulta 24 el usuario ingresa al refinador por la palabra *therapy* y se mueve por la jerarquía incorporando a la búsqueda otros términos relacionados con su interés de búsqueda, tales como *psychotherapy* y *psychoanalysis*. Algo similar ocurre con *depression* donde agrega ciertos tipos específicos de depresión.

- *Consultas con sinónimos y siglas polisémicas agregados de WordNet*

En consultas como la 1 y la 18, la inclusión de siglas de muy pocas letras como sinónimos, puede ser un problema para la recuperación. Por ejemplo, siglas como *us* o *usa* para Estados Unidos puede traer muchos documentos y no relevantes.

En la consulta 23, la cantidad de documentos resultantes aumenta debido al agregado de sinónimos a las palabras *license* y *cost*.

Un gráfico comparativo de la cantidad de documentos resultantes sin refinamiento y con refinamiento se muestra en la Figura 5.1. En la figura no se incluye la consulta 1 debido a que la cantidad de documentos resultantes de la estrategia de búsqueda obtenida luego del refinamiento semántico es excesivamente elevada y no permite utilizar una escala adecuada para visualizar el resto de las consultas. Este número excesivo de documentos recuperados se debe al gran número de sinónimos que aporta WordNet para *EEC*, incluyendo frases y siglas.

En la Figura 5.2 se muestran estos mismos resultados, pero con escala

logarítmica para no descartar y poder visualizar la consulta 1.

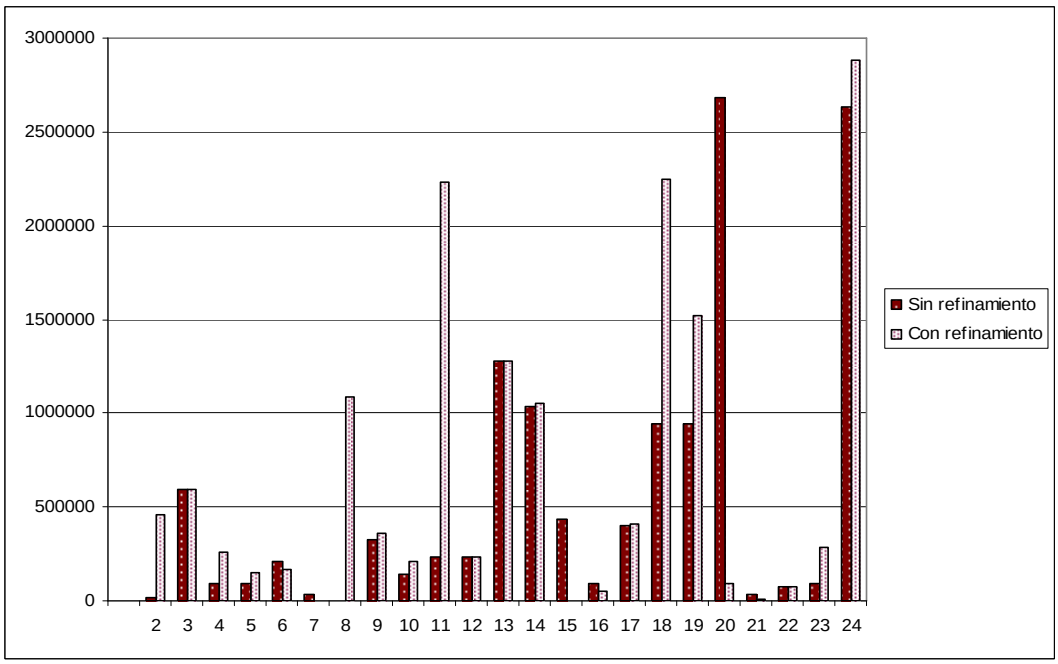


Figura 5.1: Cantidad de documentos resultantes con y sin refinamiento semántico

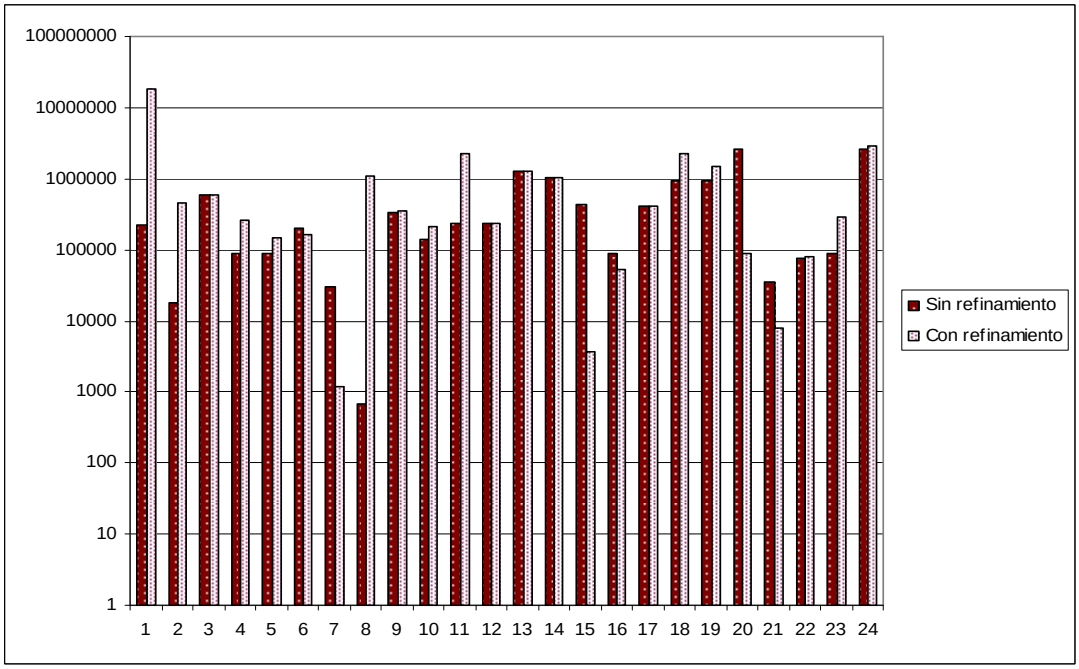


Figura 5.2: Cantidad de documentos resultantes con y sin refinamiento semántico en escala logarítmica

5.1.2. Precisión en los primeros 50 documentos

A fin de medir la precisión en los primeros 50 documentos, se midió la cantidad de documentos relevantes recuperados en estos primeros 50 documentos.

- *Consultas con nombres propios y siglas*

En consultas como la 3, donde *Mendel* es un nombre propio, y la 13, donde *GSM* es una sigla, pero que no están en WordNet, no cambian los valores obtenidos con respecto a la consulta sin refinamiento semántico en el buscador Yahoo!. La cantidad de documentos relevantes recuperados también se mantiene en consultas como la 12, donde la estrategia con refinamiento no varía de la planteada por el usuario *experto*. Algo similar se puede observar en la consulta 9 donde la cantidad de documentos relevantes recuperados con refinamiento y sin refinamiento es la misma. Esto se debe a que la estrategia resultante del refinamiento no difiere mucho de la planteada por el usuario experto.

- *Consultas con frases*

Con el refinamiento semántico, en las consultas 1, 2, 6, 7, 8, 10, 15, 16, y 21 se observa que aumenta la cantidad de documentos relevantes recuperados. Esto se debe a que en general el usuario no utiliza la búsqueda por frases en las expresiones de búsqueda sin refinar. El refinamiento semántico realiza la búsqueda por frases en el caso que el término esté en WordNet y sea compuesto.

Aún si el concepto está formado por varias palabras y no está en WordNet, el refinamiento semántico también fuerza a la construcción de la frase como ocurre por ejemplo en la consulta 7, donde busca por la frase “*Peugeot Partner”*, a pesar que el usuario no ingresó las comillas.

- *Consultas con errores ortográficos*

En consultas como la 8, la cantidad de documentos relevantes recuperados es notablemente mayor con el refinamiento semántico. Esto se debe a que la palabra

ingresada, *colli*, estaba mal escrita. En este caso, el refinamiento ofrece la posibilidad de buscar la palabra *coli*, la cual es el término ortográficamente correcto y más preciso.

- *Consultas con términos más específicos*

En la consulta 10, la cantidad de documentos relevantes recuperados con refinamiento también aumenta. Esto se debe a que el usuario ingresa por el término *lymphoma*, se mueve por la jerarquía conceptual asociada y decide reemplazar este término de partida por *"hodgkin's disease"*, que es un tipo específico de linfoma y que responde mejor a su interés, por lo cual la precisión de la respuesta es mayor. Algo similar ocurre en la consulta 20, donde el usuario al ingresar al refinador, se movió por la jerarquía del término *food* y decidió quedarse con un término más específico "*diabetic diet"*. Esto aumentó la precisión de la respuesta.

En la consulta 24, donde el usuario elige varios términos específicos a su interés de búsqueda, también se aumenta la precisión.

Las consultas 11 y 12 corresponden a un mismo interés de búsqueda y están resueltas con dos formas distintas de realizar el refinamiento semántico. En la consulta 11, el usuario ingresa tres términos: *spanish, civil* y *war*, y el refinamiento amplía cada uno de éstos con los respectivos sinónimos. En la consulta 12, el usuario decide ingresar por el término *war* y recorriendo la jerarquía conceptual baja de nivel a "*civil war"* y dentro de éste, baja nuevamente de nivel para optar por un tipo particular de guerra civil: *"spanish civil war"*. En la consulta 11, la cantidad de documentos relevantes recuperados con refinamiento semántico es mucho menor que a la obtenida con refinamiento semántico en la consulta 12. Esto se debe a que en la consulta 11, el usuario ingresó como conceptos para el refinamiento los adjetivos *spanish* y *civil*, que en realidad no son conceptos sino adjetivos calificativos de *war*. Lo correcto es, en casos como éste, realizar una estrategia como la de la consulta 12, donde se ingresa por el sustantivo principal y se eligen, recorriendo la jerarquía conceptual, como términos específicos los sustantivos adjetivados. De esta manera se aumenta la precisión.

En las consultas 4 y 5, el interés del usuario es buscar páginas que hablen sobre ganadores del premio nobel de medicina. Aquí ocurre un caso similar al visto con las consultas 11 y 12. Es distinto buscar a *Nobel* como persona, como en la consulta 4

donde WordNet ofrece *"Alfred Nobel"* y *"Alfred Bernhard Nobel"* como sinónimos, o buscar *"nobel prize"*, como en la consulta 5, donde *"nobel prize"* es un tipo de *prize*, y por lo tanto es más adecuada para el interés del usuario de este ejemplo.

- *Consultas con términos con pocos o ningún sinónimo en WordNet*

En las consultas 14, 17, y 22, no se observan variaciones importantes de cantidad de documentos relevantes con y sin refinamiento. Las expresiones de búsqueda con refinamiento tienen pocos sinónimos agregados, y los términos utilizados por los usuarios no corresponden a nombres propios o siglas.

En la Figura 5.3. se grafica la precisión en los primeros 50 documentos sin refinamiento y con refinamiento semántico. Esta precisión se calcula dividiendo el número de documentos relevantes recuperados en los primeros 50 documentos dividido 50.

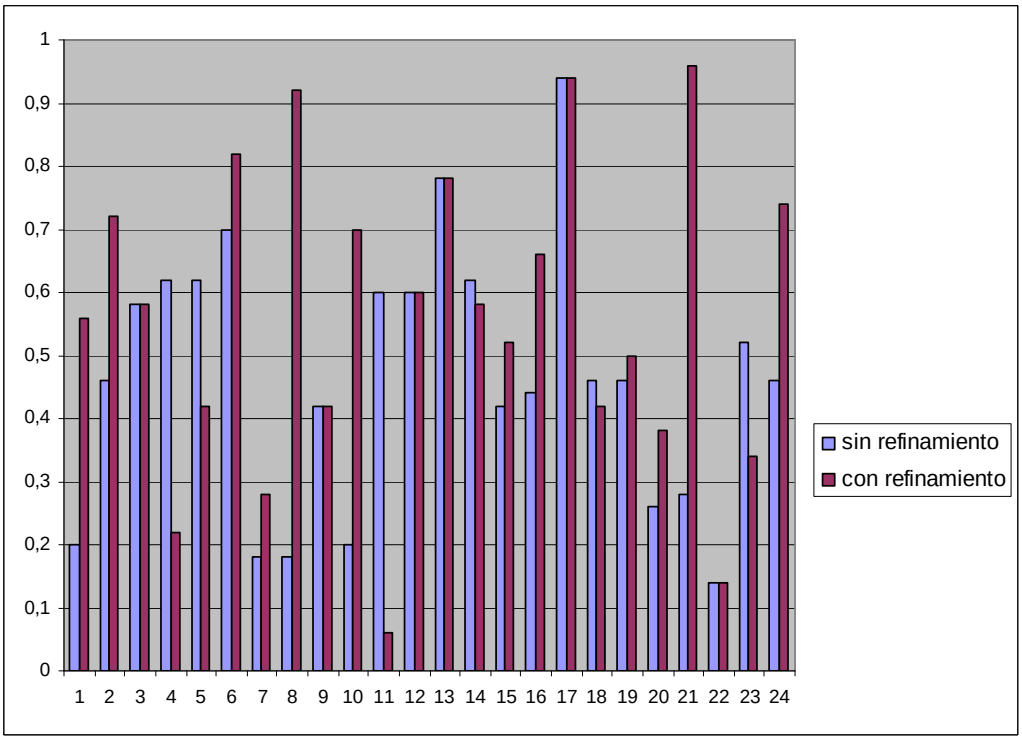


Figura 5.3: Precisión en los primeros 50 documentos, con y sin refinamiento semántico

5.1.3. Conclusiones de la experimentación con WordNet

El objetivo de las experiencias realizadas es evaluar la utilización del recurso lingüístico de cobertura general WordNet para la preparación de la estrategia de búsqueda para la recuperación de información en la Web, en consultas de interés general. De los resultados obtenidos se puede observar que:

- En general, el usuario no utiliza la búsqueda por frases. Por ejemplo, Gabriel García Márquez son tres palabras que forman parte de un solo concepto y debería buscarse como una unidad: “Gabriel García Márquez”. El refinamiento semántico genera automáticamente frases a partir de conceptos formados por más de una palabra, ya sea que estos conceptos estén en WordNet o no. El uso de frases en la estrategia de búsqueda aumenta la precisión de la recuperación, y disminuye la cantidad de documentos recuperados.
- En el caso de nombres propios que no están en WordNet no varía la precisión con respecto a la búsqueda sin refinamiento, excepto que estos nombres propios sean frases, en cuyo caso la precisión mejora.
- La estrategia generada con refinamiento semántico no difiere mucho de la planteada por un usuario experto. Por lo tanto, los resultados de la búsqueda con refinamiento son bastante similares a los resultados sin refinamiento.
- La estrategia generada con refinamiento semántico mejora la precisión en el caso de usuarios inexpertos o medios.
- En general, el usuario no ingresó términos con errores ortográficos, pero en la única consulta (Consulta 8) donde ingresó con errores ortográficos, la corrección ortográfica realizada por el refinamiento aumentó la cantidad de documentos recuperados y la precisión de los mismos.
- Mediante el refinamiento semántico se permite la navegación por una jerarquía conceptual, donde al poder seleccionar el usuario términos más específicos aumenta la precisión.
- La utilización de sustantivos adjetivados, como por ejemplo “*spanish civil*

war”, como un solo concepto para el refinamiento semántico, aumenta la precisión y disminuye la cantidad de documentos recuperados. Los sustantivos adjetivados se obtienen moviéndose por la jerarquía conceptual a partir del sustantivo, en este ejemplo *war*.

- Analizado el número de conceptos utilizados en cada consulta, en aquellas que involucran más de un concepto, el promedio de la cantidad de documentos recuperados aumentó luego del refinamiento semántico. En el caso de consultas que involucran un solo concepto, si el refinamiento consiste en sólo agregar sinónimos, aumenta la cantidad de documentos recuperados. Pero, si el refinamiento consiste en cambiar el concepto inicial por uno más específico, la cantidad de documentos recuperados disminuye. Los resultados se presentan en la Tabla 5.2. y en la Figura 5.4.

	Sin refinamiento	Con refinamiento
1 concepto	669930,00	342760,00
2 conceptos	570218,18	2298989,09
3 conceptos	359928,57	985142,86

Tabla 5.2: Promedio cantidad de documentos recuperados según cantidad de conceptos utilizados

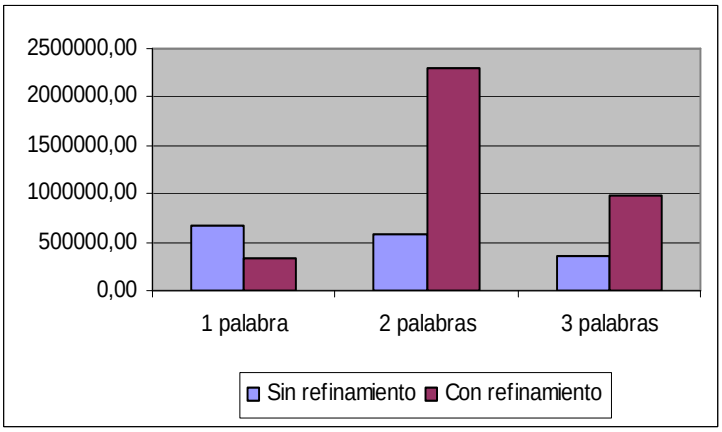


Figura 5.4: Promedio cantidad de documentos recuperados según cantidad de conceptos utilizados

- Analizado el número de conceptos utilizados en cada consulta, el promedio de la precisión aumentó luego del refinamiento semántico para consultas que involucran uno o dos conceptos. Para las consultas que involucran más de dos conceptos, hay dos posibles causas de la disminución de la precisión en la experiencia realizada. La primera causa es la no utilización de sustantivos adjetivados; este es el caso donde el usuario ingresó *spanish*, *civil* y *war* como tres conceptos en lugar de considerarlo un solo concepto, como es la forma correcta. La segunda causa es el agregado por parte del refinador de siglas cortas como sinónimos, por ejemplo, se agregan *US*, *USA* para *United States*. Los resultados se presentan en la Tabla 5.3. y en la Figura 5.5.

	Sin refinamiento	Con refinamiento
1 concepto	0,41	0,61
2 conceptos	0,47	0,63
3 conceptos	0,50	0,38

Tabla 5.3: Promedio de precisión según cantidad de conceptos utilizados

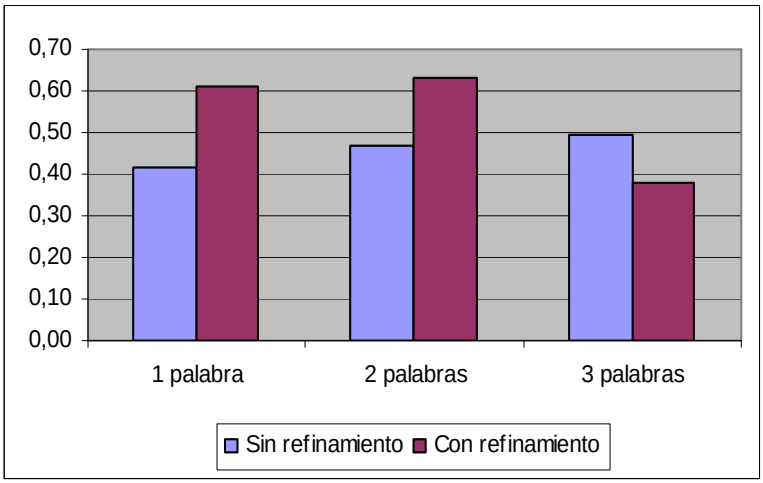


Figura 5.5: Promedio de precisión según cantidad de conceptos utilizados

Finalmente, se promediaron la cantidad de documentos recuperados y la precisión sobre los primeros 50 resultados sin y con refinamiento semántico. Los resultados se muestran en la Tabla 5.4.

De las 24 consultas realizadas, se descartó la consulta 1 debido a la gran cantidad de documentos resultantes de la estrategia de búsqueda obtenida luego del refinamiento semántico. Este número excesivo de documentos recuperados se debe al gran número de sinónimos que aporta WordNet para *EEC*, incluyendo frases y siglas.

	Recuperados	Precisión
Sin refinamiento	533811,67	0,46
Con refinamiento	651726,67	0,55
	22,09 %	19,03 %

Tabla 5.4: Promedios de cantidad de documentos recuperados y precisión sobre los primeros 50 resultados

De los promedios se observa que el refinamiento semántico mejora la cantidad de documentos recuperados en un 22,09 % y mejora la precisión en un 19,03 %. Esto muestra que la propuesta de refinamiento semántico presentada mejora la recuperación de información de la web al utilizar WordNet como recurso lingüístico de cobertura general para la preparación de la estrategia de búsqueda. Estos resultados no difieren mucho de los presentados por [Sangoi Pizzato et al., 2003] en un trabajo similar donde la expansión de la consulta se basa en tesauros en lugar de WordNet como recurso lingüístico.

Los resultados de esta experiencia fueron aceptados para su publicación en los Proceedings del LatinAmerican Web Congress [Deco et al., 2005].

5.2. Experimentación en un dominio específico

Wordnet es un recurso lingüístico de cobertura general, que mejora la recuperación de información en consultas de interés general, como se mostró en la experimentación realizada de la Sección 5.1. Es de interés en esta tesis evaluar, también, el refinamiento semántico en consultas realizadas en dominios específicos del conocimiento, utilizando recursos lingüísticos especializados. Esta experimentación se decide llevarla a cabo en el área Salud, debido a que la autora de esta tesis, participó durante varios años en proyectos de investigación y desarrollo en dicha área. En particular, esta experimentación se realizó en el marco del Proyecto de Investigación "Tecnologías Middleware e Internet: búsqueda asistida de evidencia clínica en medicina" [Plüss, 2004-2006].

Para la experimentación en este dominio específico, se utilizó el recurso lingüístico MeSH [MeSH], por ser uno de los más utilizados en el área salud. MeSH, es el vocabulario controlado o tesauro de la base de datos Medline [Medline]. Este recurso presenta los conceptos en una estructura jerárquica, distribuidos desde los más genéricos a los más específicos. Medline es una de las bases de datos bibliográficas que componen Medlars (Medical Literature Analysis Retrieval System), elaborada por la Biblioteca Nacional de Medicina de Estados Unidos y está disponible online.

Para la experimentación se realizaron 25 consultas. Para esto se citaron médicos de distintas especialidades, con y sin conocimientos en búsqueda en la Web, y se les solicitó que realizaran búsquedas. En primer lugar, los profesionales realizaron la consulta como lo hacen habitualmente y se registró la cantidad total de documentos devueltos por el buscador y la cantidad de páginas relevantes en los primeros 10, 20 y 50 enlaces obtenidos. Luego, se realizó cada búsqueda utilizando el refinamiento semántico de la consulta. En este caso también se registraron la cantidad total de documentos devueltos por el buscador y la cantidad de páginas relevantes en los primeros 10, 20 y 50 enlaces obtenidos a fin de medir la precisión en estos documentos. En la Tabla 5.5 se muestran las consultas realizadas.

		Búsqueda sin refinamiento en Yahoo!					Búsqueda con refinamiento en Yahoo!					
Nro. consulta	Descripción	Expresión búsqueda	Cantidad docs. resultantes	Cant. docs. relevantes en los primeros docs.			Términos para el refinamiento semántico	Estrategia de búsqueda	Cantidad docs. resultantes	Cant. docs. relevantes en los primeros docs.			Nivel de usuario
				10	20	50				10	20	50	
1	Uso de lincomicina en embarazadas	Uso de lincomicina en el embarazo	94	3	3	5	- Lincomicina - Embarazo	(“Lincomicina” OR “Lincomycin”) AND (“Embarazas” OR “Neophroparty”)	103	3	8	14	Inexperto
2	Tratamiento de interferón pegilado en Hepatitis C	Tratamiento de interferón pegilado en Hepatitis C	1200	4	8	19	- Interferón pegilado - Hepatitis C	"Tratamiento" AND "Hepatitis C" AND "Interferón"	902	6	11	26	Inexperto
3	Litiasis Renal	Litiasis Renal	37900	3	8	13	Litiasis Renal	"Litiasis Renal" OR "Lithiases Renal" OR "Calculosis Renal" OR "Calculoses Renal"	16600	7	11	28	Inexperto
4	Colitis Ulcerosa	"Colitis Ulcerosa"	493000	4	6	13	Colitis Ulcerosa	"Colitis Ulcerosa" OR "Colitis ulcerative" OR "Ulcerative colitis"	3640000	9	16	40	Medio
5	Sindrome de Sjögren	"Sindrome de Sjögren"	15200	3	6	12	Sindrome de Sjögren	"Sindrome de Sjogren" OR "Sjogren Syndrome" OR "Sjogrens Syndrome" OR "Syndrome Sjogren's" OR "Sicca Syndrome" OR "Syndrome, Sicca"	1030000	6	11	31	Medio
6	Enfermedad de Crhon	"Enfermedad de Crhon"	77	3	5	7	Enfermedad de Crhon	"Enfermedad de Crhon" OR "Crohn Disease" OR "Enfermedad de Crohn"	6290000	5	14	31	Medio
7	Polimialgia reumatica	Polimialgia reumatica	2010	4	6	11	Polimialgia reumatica	"Polimialgia reumatica" OR "Polymyalgia rheumatica"	322000	2	7	28	Inexperto
8	Enfermedad de Munchausen	"Enfermedad de Munchausen" OR "Sindrome de Munchausen"	608	5	7	7	Sindrome de Munchausen	"Sindrome de Munchausen" OR "Syndrome Munchausen" OR "Syndrome Hospital-Addiction" OR "Munchhausen Syndrome" OR "Hospital-Addiction Syndrome"	20700	5	7	17	Medio
9	Hepatitis aguda tratada con Paracetamol	"Hepatitis aguda" "paracetamol"	131	5	7	12	Hepatitis aguda, "paracetamol"	("Hepatitis aguda" AND "paracetamol") OR ("Hepatitis Chronic" AND Paracetamol)	216	5	13	25	Medio
10	Colotis	Colotis	1070	4	8	18	Colotis	"Pseudomembranous Colitis" OR	121000	8	17	27	Inexperto

		Búsqueda sin refinamiento en Yahoo!					Búsqueda con refinamiento en Yahoo!						
Nro. consulta	Descripción	Expresión búsqueda	Cantidad docs. resultantes	Cant. docs. relevantes en los primeros docs.			Términos para el refinamiento semántico	Estrategia de búsqueda	Cantidad docs. resultantes	Cant. docs. relevantes en los primeros docs.			Nivel de usuario
	Pseudomembranosa	Pseudomembranosa					Pseudomembranosa	"Colitis pseudomembranous" OR "Pseudomembranous colitis"					
11	Purpura Trombocitopénica Idipática	"Purpura Trombocitopénica Idiopática"	1050	7	15	22	Purpura Trombocitopénica Idipática	"Purpura Thrombocytopenic Idiopathic" OR "Purpura Trombocitopénica Idiopática "	2070	7	14	25	Medio
12	Tratamiento con insulina en la embarazada diabética	"Tratamiento" "Insulina" "Embarazadas Diabéticas"	74	4	8	15	Tratamiento - Insulina - Embarazadas Diabéticas	("Tratamiento" AND "Insulina" AND "Embarazadas Diabéticas") OR ("Treatment" AND "Insulin" AND "Neophropathy Diabetic")	58	4	6	18	Experto
13	La tartamudez	Tartamudez	49900	3	8	18	Tartamudez	tartamudez OR disfluencies OR disfluencias	102000	4	6	15	Inexperto
14	Rechazo del transplante renal	"Transplante Renal" AND "rechazo"	429	1	4	21	Transplante Renal - Rechazo	("transplante renal" AND "rechazo cronico") OR ("transplant renal" AND ("transplante renal" AND "rechazo cronico") OR ("transplant renal" AND ("Renal rejection" OR "rejection Renal " OR "Renal Failure" OR "Failure Renal ")))	871	5	12	27	Inexperto
15	"Cómo evoluciona el tratamiento del HPV con una cirugía"	Tratamientos HPV	69400	8	14	34	"HPV" and (cricirugía or crioterapia or electrocoagulación)	HPV AND (cricirugía OR crioterapia OR electrocoagulación)	19	9	16	16	Medio
16	"Tratamientos para "borrar" las manchas solares de la piel	"skin sunspots" treatment	2560	2	8	11	"skin sunspots" and treatment	("manchas solares" tratamiento) OR ("skin sunspots" treatment)	392	7	10	12	Experto
17	"Eliminar las manchas solares de la piel mediante la microdermoabrasión"	"skin sunspots" microdermabrasion	18	5	5	5	"skin sunspots" and microdermabrasion	("manchas solares" microdermoabrasión) OR ("skin sunspots" microdermabrasion)	48	7	11	11	Experto
18	"Vacuna contra el	"skin cancer	2030000	10	12	19	"skin cancer	(vacuna "cáncer piel") OR ("skin cancer"	1130	9	16	26	Experto

		Búsqueda sin refinamiento en Yahoo!					Búsqueda con refinamiento en Yahoo!					
Nro. consulta	Descripción	Expresión búsqueda	Cantidad docs. resultantes	Cant. docs. relevantes en los primeros docs.			Términos para el refinamiento semántico	Estrategia de búsqueda	Cantidad docs. resultantes	Cant. docs. relevantes en los primeros docs.			Nivel de usuario
	cáncer de piel"	vaccine"					vaccine"	vaccine)					
19	"Formas de prevención de la diabetes"	prevención de la diabetes	61100	9	17	42	prevención, diabetes	(prevención OR prevention) AND diabetes	36.400.000	10	20	49	Medio
20	"Diferentes manifestaciones de hipotiroidismo en la niñez"	hipotiroidismo en la niñez	704	9	15	35	hipotiroidismo, niñez	hipotiroidismo AND (childhood OR niñez)	10700	10	18	42	Medio
21	"Infarto agudo de miocardio"	infarto agudo miocardio	407000	8	16	33	infarto, agudo, miocardio	"infarto agudo miocardio" OR (infarct, myocardial) OR (infarction, Miocardial)	58600	8	17	42	Medio
22	"Leucemia mieloide crónica"	"leucemia mieloide crónica"	127000	10	19	40	leucemia, mieloide, crónica	("leucemia mieloide crónica") OR (Leukemia, Myeloid, Chronic)	851	9	18	45	Medio
23	"Neumonía de Friedländer"	Neumonía de Friedländer	166	10	17	32	neumonía, Friedländer						Inexperto
24	"Peritonitis Aguda"	peritonitis aguda	96600	9	16	28	peritonitis, aguda	("peritonitis aguda") OR ("acute peritonitis")	15800	8	15	30	Inexperto
25	"Úlcera gástrica"	úlcera gástrica	315000	8	14	24	úlcera, gástrica	("úlcera gástrica") OR ("gastric ulcer") OR ("stomach ulcer")	30300	10	18	38	Inexperto

Tabla 5.5.: Consultas realizadas en un dominio específico

En la Figura 5.6 se grafica la cantidad total de documentos resultantes que se obtuvo en la búsqueda sin refinamiento y en la búsqueda con refinamiento semántico. Como se puede apreciar, en la mayoría de los casos, la búsqueda resultante de la estrategia obtenida con refinamiento, arrojó una cantidad mayor de documentos que la obtenida a partir de la estrategia original escrita por el usuario.

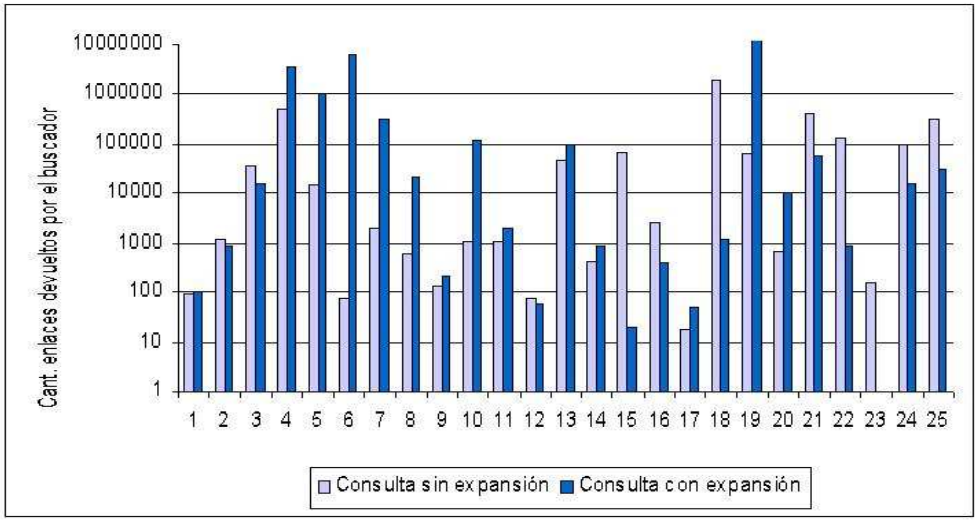


Figura 5.6: Cantidad de enlaces resultantes obtenidos en la consulta sin refinamiento y con refinamiento

En las Figuras 5.7, 5.8 y 5.9 se muestran la precisión en los primeros 10, 20 y 50 documentos respectivamente, para la búsqueda original y para la búsqueda con refinamiento.

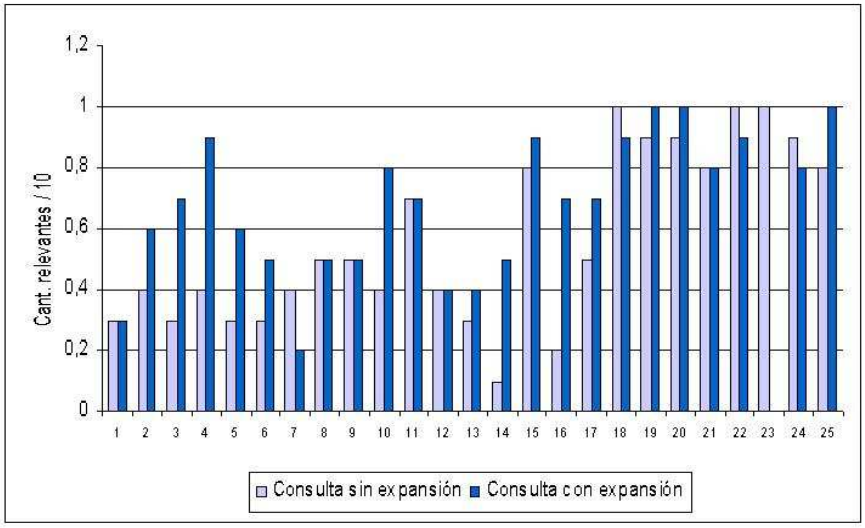


Figura 5.7: Precisión en los primeros 10 documentos

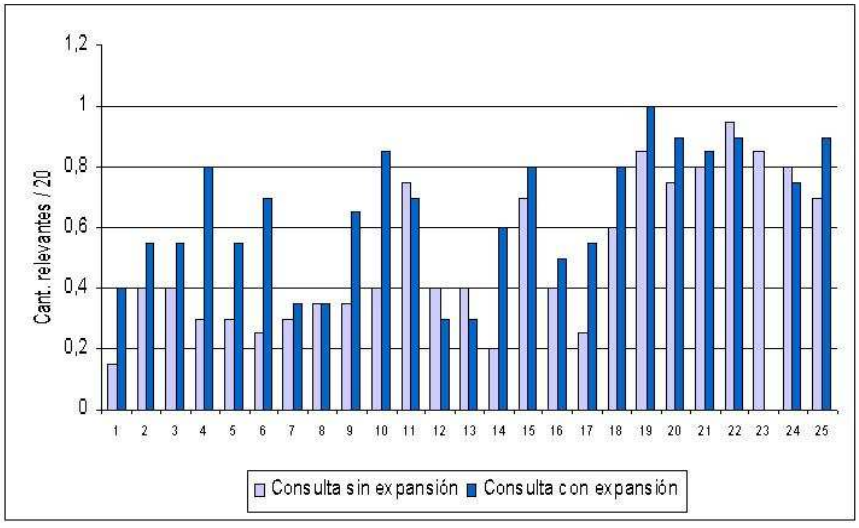


Figura 5.8: Precisión en los primeros 20 documentos

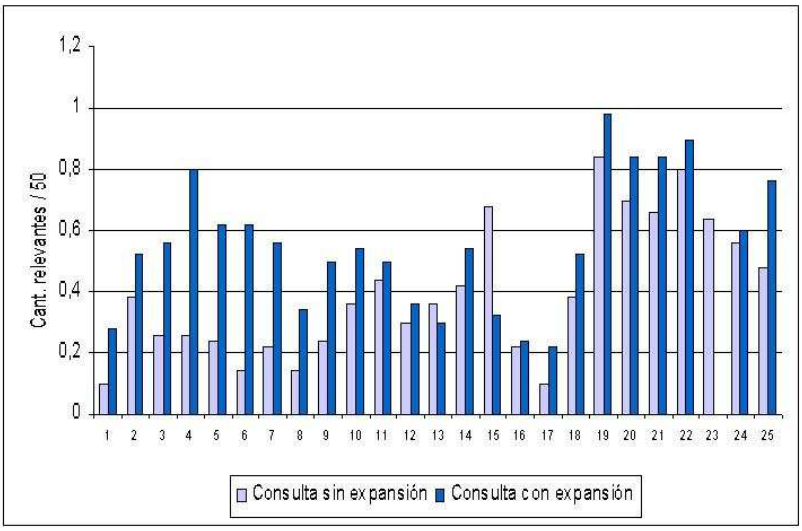


Figura 5.9: Precisión en los primeros 50 documentos

En los tres gráficos se puede apreciar que la utilización del modelo propuesto arrojó mejores resultados que la búsqueda tradicional. De esta manera, no sólo se mejora la cantidad de resultados sino la calidad de los mismos.

5.2.1. Discusión sobre trabajos relacionados en dominios específicos

Existen algunos trabajos relacionados para mejorar la búsqueda de información en dominios específicos del conocimiento. A continuación se describen algunos correspondientes al área salud, ya que esta área fue una de las primeras en generar profusa documentación en formato electrónico, y por esto es pionera en proponer buenas interfaces de búsqueda para los usuarios médicos. Uno de los primeros proyectos fue CITE (Current Information Transfer in English) [Doszkocs, 1982], que es una interfaz a Medline donde el usuario ingresa las consultas en lenguaje natural. Utiliza ponderación y realimentación por relevancia para la expansión de la consulta, agregando a la consulta original los términos MeSH de los documentos recuperados que el usuario marcó como relevantes. El proyecto RBR-EVI [French et al., 2001] expande la consulta agregando los tres términos MeSH mejor rankeados. En algunos sistemas la consulta se construye utilizando

directamente la estructura MeSH, como en MenUSE [Pollitt, 1988], donde el usuario ingresa un término y desciende en el árbol MeSH correspondiente al mismo hasta alcanzar el significado específico de interés. El mismo proceso se repite para cada uno de los términos de la consulta. Meva (Medline Evaluator) [Meva] es un sistema de post-procesamiento de Medline que permite al usuario refinar la estrategia sobre los documentos resultantes de una búsqueda aplicando distribuciones de frecuencia y tablas de contingencia para detectar los términos que co-ocurren en los mismos para utilizarlos en la expansión de la consulta. HONselect [HONselect] es un catálogo de términos médicos y un integrador de búsquedas. Permite búsquedas en cinco idiomas: Inglés, Francés, Alemán, Español y Portugués, ingresando la consulta en el lenguaje elegido por el usuario y separando los resultados de cada idioma. Para la expansión de consulta usa MeSH, mostrando la jerarquía original en inglés con los términos traducidos al idioma designado. Se recuperan resultados de diversas fuentes: Servicio de Noticias de Yahoo, artículos médicos de Medline a través de PubMed, e información desde bases de datos de la Fundación HON sobre recursos web, multimedia, conferencias y eventos. BIREME [Bireme] ofrece en su sitio web un servicio que emplea DeCS, que es la traducción al español y al portugués del tesauro MeSH, y que realiza búsqueda y recuperación en las bases de datos Lilacs, Medline y otras. El usuario puede elegir entre los tres idiomas: inglés, español y portugués, y expandir la consulta navegando la estructura jerárquica de DeCS o refinarla agregando términos de la jerarquía de calificadores, pero no automatiza la expansión.

La mayoría de los proyectos mencionados utilizan MeSH como recurso lingüístico, al igual que la propuesta presentada aquí. Los proyectos mencionados, amplían la búsqueda en una sola dirección: algunos expanden los conceptos semánticamente pero sin mostrar la jerarquía de MeSH; otros permiten seleccionar un concepto más específico de la jerarquía de MeSH pero sin realizar expansión por sinónimos. Además, ninguno realiza una corrección ortográfica del término ingresado.

En esta tesis se propone ampliar la cantidad de documentos recuperados expandiendo el concepto semánticamente, previa verificación ortográfica del concepto a buscar. Además, se propone mejorar la precisión a través de una

interacción mínima del usuario, para la desambiguación del concepto a fin de presentar la jerarquía de conceptos relacionada con la acepción de interés, para que el usuario pueda incorporarlos a su consulta. El sistema de búsqueda semiautomatizada que se presenta tiene la ventaja de su simplicidad y su bajo costo de implementación y mantenimiento. Además, el sistema guía al usuario en todo el proceso, lo que es útil para los usuarios inexpertos de esta área del conocimiento en la búsqueda de información.

5.2.2. Conclusiones de la experimentación en un dominio específico

El objetivo de estas experiencias es evaluar la utilización de un recurso lingüístico especializado para la preparación de la estrategia de búsqueda en consultas en dominios específicos, en particular en el área salud. Para las experiencias se seleccionó el recurso MeSH del dominio médico. De los resultados obtenidos se puede observar que:

- En general, el usuario no utiliza la búsqueda por frases. El refinamiento semántico genera automáticamente frases a partir de conceptos formados por más de una palabra. El uso de frases en la estrategia de búsqueda aumenta la precisión de la recuperación, y disminuye la cantidad de documentos recuperados.
- En el caso de nombres propios, tales como nombres de enfermedades, se observa que en general estos nombres están en el recurso MeSH. En cambio, en el recurso general WordNet estos nombres en general no están. Por esto, en búsquedas específicas es preferible el uso de un recurso específico en lugar de uno de cobertura general.
- La estrategia generada con refinamiento semántico no difiere mucho de la planteada por un usuario experto.
- La estrategia generada con refinamiento semántico mejora la precisión en el caso de usuarios inexpertos o medios.
- Mediante el refinamiento semántico se permite la navegación por una

jerarquía conceptual, donde al poder seleccionar el usuario términos más específicos aumenta la precisión.

En la Tabla 5.6 se muestran los porcentajes promedios de la precisión en los primeros 10, 20 y 50 documentos, tanto para la consulta original como para la consulta con refinamiento semántico.

	Consultas sin refinamiento	Consultas con refinamiento	Mejora de la Precisión
Primeros 10	0,564	0,652	15,60 %
Primeros 20	0,504	0,624	23,81 %
Primeros 50	0,397	0,530	33,50 %

Tabla 5.6: Porcentajes promedio de precisión

Se observa que al considerar los 10 primeros documentos se mejoró la precisión en un 15,60 %. En el caso de considerar los 20 primeros documentos se mejoró en un 23,81 %. Y en los 50 primeros documentos la mejora de la precisión es de un 33,50 %.

Los resultados finales de esta experiencia se publicaron en “Un Sistema para Mejorar la Recuperación de Información Médica en la Web mediante la Expansión Semiautomática de la Consulta” en la Revista Española de Informática y Salud [Deco et al., 2006]. Propuestas previas y resultados parciales fueron presentados en [Deco et al., 2005b] y [Bender et al., 2006b].

5.3. Conclusiones de la experimentación

De los promedios de las dos experiencias anteriores se observa que el refinamiento semántico mejora la precisión en los primeros 50 documentos en un 19,03% en un dominio general del conocimiento utilizando recursos generales y en un 33,50% en un dominio específico utilizando recursos especializados.

Esto muestra que la propuesta de refinamiento semántico presentada mejora la recuperación de información de la web al utilizar recursos lingüísticos para la preparación de la estrategia de búsqueda. Además estos resultados mejoran en un dominio específico del conocimiento si se utiliza un recurso especializado.

Capítulo 6. Refinamiento semántico aplicado a la búsqueda en un sitio web

En la actualidad, existen numerosos sitios web que manejan un caudal de información considerable. Estos sitios suelen tener un buscador para permitirle al usuario localizar y acceder a la información que necesita en forma directa y más rápida, evitando tener que navegar todo el sitio en cuestión. Al igual que la búsqueda en la web, en la búsqueda dentro de un determinado sitio, muchas veces cuando se busca información a través de un buscador, no se obtienen los resultados esperados. Estas búsquedas infructuosas pueden ocurrir debido a un conjunto de factores que inciden negativamente en los resultados de las búsquedas. En este tipo de escenarios se presentan también problemas de sinonimia y polisemia de términos.

En este capítulo se plantea el objetivo siguiente: dado un sitio web, mejorar las prestaciones del motor de búsqueda de dicho sitio mediante el refinamiento semántico propuesto en esta tesis, y por lo tanto mejorar la efectividad de las búsquedas. Para esto, se adaptó la arquitectura propuesta en la Sección 4.1., para la búsqueda en un sitio web y no en toda la web. Si bien la propuesta es aplicable a cualquier dominio del conocimiento, la experimentación se realizó sobre un sitio de Turismo.

6.1. Incorporación del Refinamiento Semántico a un buscador de un sitio web

Un buscador tiene tres componentes principales: un *crawler*, un *indexador* y un *motor de búsqueda*. El crawler recorre la colección y crea un repositorio local con el contenido de las páginas visitadas. El indexador procesa dicho repositorio y genera un índice invertido del mismo. Las consultas ingresadas son enviadas al motor de búsqueda. Este último consulta al índice y retorna los resultados de las búsquedas al usuario.

La incorporación del refinador semántico en la arquitectura general de un buscador se muestra en la Figura 6.1. La arquitectura está compuesta por un motor de búsqueda web (buscador), un índice invertido, una ontología y una colección de páginas web etiquetadas con metadatos.

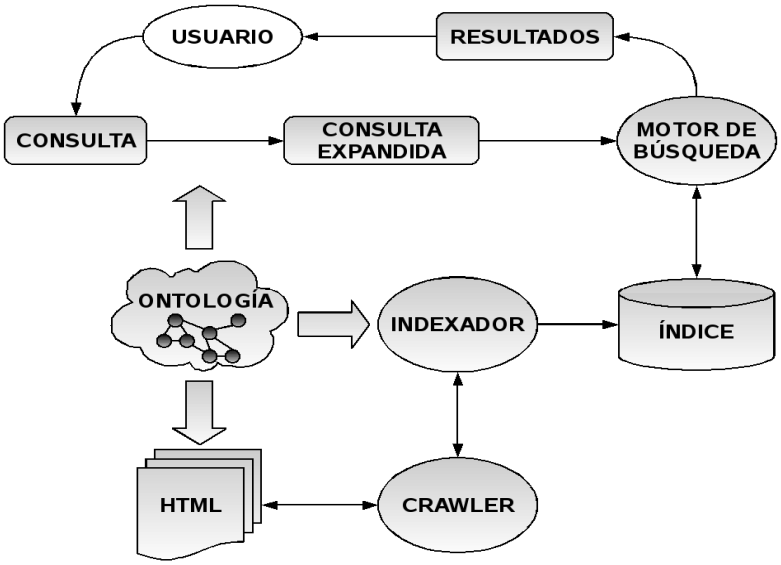


Figura 6.1: Arquitectura del buscador con refinamiento semántico

En esta incorporación un elemento clave es el uso de una *ontología de dominio* que es utilizada como recurso lingüístico para el refinamiento. La ontología de dominio se utiliza además para etiquetar las páginas del sitio. La consulta ingresada por un usuario, en lenguaje natural, es expandida con información extraída de la ontología, obteniendo la estrategia de búsqueda correspondiente que es enviada al motor de búsqueda.

La ontología utilizada es específica del dominio de aplicación del sitio web. Para la experimentación se utiliza un sitio web relativo al turismo en Argentina.

El propósito del etiquetado de las páginas del sitio es utilizar términos de un vocabulario controlado con el fin de reducir la sinonimia y polisemia de términos. El etiquetado permite insertar información adicional (metadatos) en los documentos de la colección, de modo que queden caracterizados mediante algún criterio y por lo tanto sean más fáciles de recuperar. Para esto se clasifican los documentos de la colección utilizando la ontología, dado que ésta contiene la taxonomía del dominio. Para la experimentación de este capítulo, la clasificación de los documentos se realiza en forma manual, ya que es difícil automatizar este proceso, lo cual no es objetivo de esta tesis. La dificultad radica en que los documentos deben caracterizarse por su contenido y no por los términos que ocurren en él. Una vez

clasificado el documento, se lo etiqueta con los metadatos, que consisten en la lista con los nombres de las clases, subclases y posiblemente individuales de la ontología. Las etiquetas se indexan bajo un campo común, llamado *category*, lo cual permite posteriormente consultar al índice por páginas que contengan cierto valor en dicho campo. Esta información es utilizada en la expansión de la consulta, tanto para la recuperación de páginas como también para incrementar la relevancia de las mismas.

Para tratar la sinonimia de términos se utiliza un vocabulario controlado. Este se construye de modo que sus términos sean los nombres de las clases de la ontología, ya que designan los conceptos del dominio. Para esto, se sustituye cada término que pueda asociarse a un concepto de la ontología, por el término del vocabulario controlado con el cual está designado dicho concepto. Para el caso de las consultas de usuario, la sustitución es realizada en la expansión de consultas, mientras que para el caso de los documentos durante la indexación, donde se indexa el término designado. Por ejemplo, dada la clase Alojamiento, que determina el término del vocabulario controlado *alojamiento,* podría estar anotada con las siguientes palabras: albergue, hospedaje y aposento, como sinónimos de alojamiento.

La estrategia de expansión de consultas consta de dos pasos. En el primero, los términos ingresados por el usuario en la consulta son sustituidos, siempre que sea posible, por los términos del vocabulario controlado. En el segundo paso, se agregan términos relacionados, generados a partir de consultas a, y/o inferencias en, la ontología. La función de estos términos relacionados es recuperar páginas que traten acerca de los conceptos asociados con los términos de la consulta y mejorar la precisión. Para ello, se realizan búsquedas sobre las etiquetas de los documentos, ya que éstas caracterizan a un documento por el concepto del que tratan.

El diseño de la ontología se realizó con el editor Protégé-OWL[25] versión 3.3.1. Para definir la taxonomía del dominio Turismo, se identificaron los conceptos, y a cada uno de estos se los representó por medio de una clase. Estas clases se organizaron en forma jerárquica, según se muestra en la Figura 6.2. En esta ontología se definieron disjuntas entre sí a las clases hermanas, con lo cual se asegura que un individual pueda pertenecer a sólo una de ellas. Por ejemplo, las clases *Movilidad*, *Alojamiento* y *Gastronomía* son clases hermanas al ser subclases directas de la clase *Servicio*; y se las hace disjuntas entre sí.

En la Figura 6.3. se muestra el gráfico de clases, propiedades y restricciones de la ontología Turismo. Las clases están representadas por rectángulos y las propiedades por flechas. Un asterisco al lado del nombre de una propiedad representa cardinalidad múltiple.

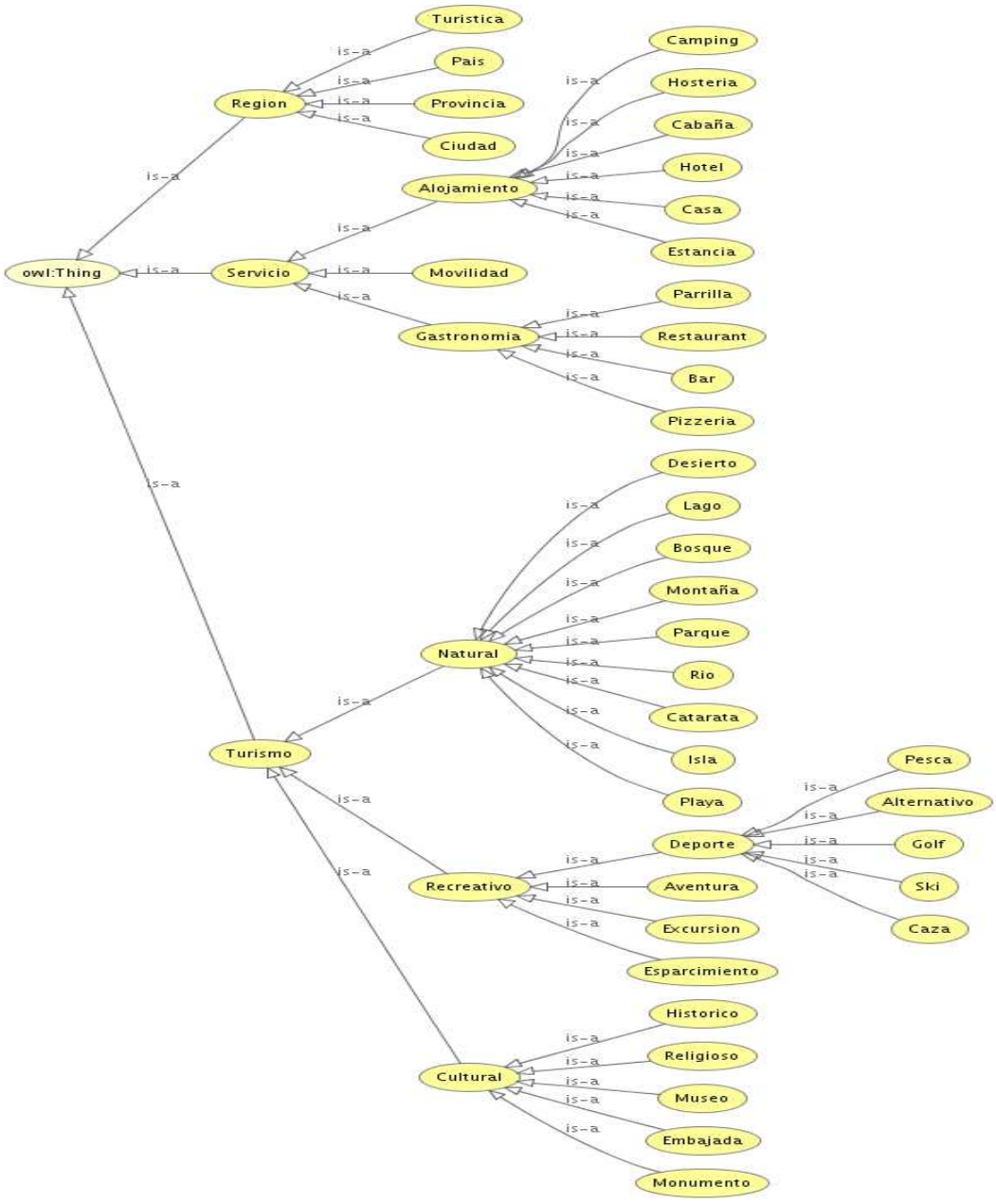


Figura 6.2: Ontología Turismo

[25] The Protégé Ontology Editor and Knowledge Acquisition System: http://protege.stanford.edu/

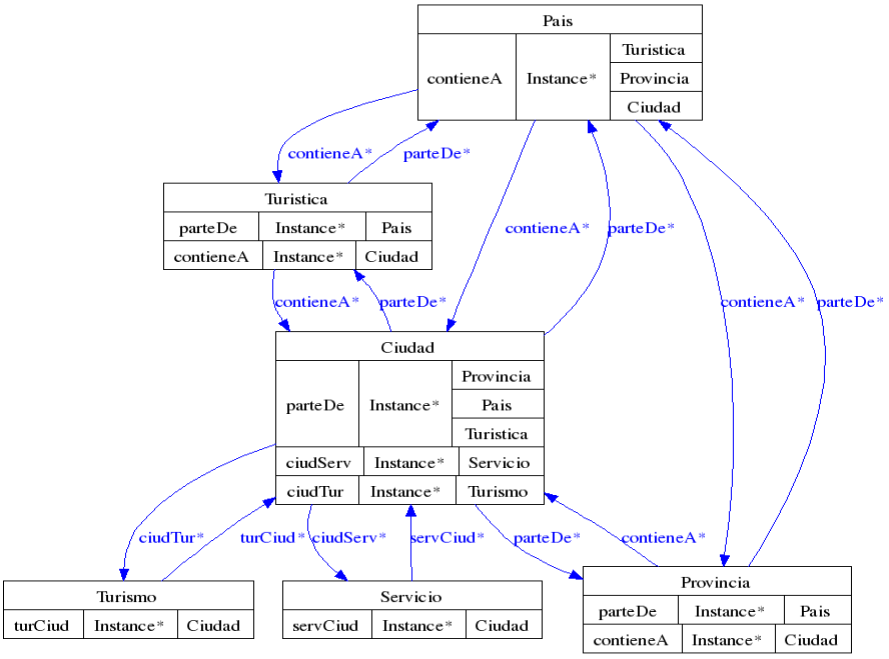


Figura 6.3. Clases, propiedades y restricciones

Los individuales se utilizan para clasificar en forma más precisa a ciertos documentos de la colección y posteriormente etiquetarlos en base a esta clasificación. Además, combinados con la expansión de consultas, permiten una mejor precisión en los resultados de la consulta. Para el prototipo se insertaron 450 individuales.

Cuando un usuario ingresa una consulta, comienza el proceso de refinamiento semántico. En primer lugar se chequea la ortografía y se realiza el stemming. Luego, se agregan términos adicionales a partir de la ontología. Estos términos adicionales se utilizan para consultar al campo *category* del índice invertido y tienen por objetivo recuperar aquellos documentos que traten sobre los conceptos asociados a los términos de consulta ingresados y aumentar la cantidad y la relevancia de los documentos recuperados.

Los términos adicionales se obtienen a partir de los términos de la consulta que puedan identificarse con el nombre de una clase o individual de la ontología, y la aplicación de un algoritmo que tiene por objetivo obtener información adicional a partir de los términos de las consultas. Para ello, puede realizar razonamientos sobre la ontología, considerando los siguientes casos:

- Si el único término identificado es una clase o un individual: retorna el nombre

de la clase o el nombre del individual respectivamente.

- Si los términos identificados son dos clases (clase1 y clase2 en ese orden):
  - Si una clase (clase1) es subclase de la otra (clase2), retorna la primera (clase1), ya que es más específica que la otra, y por lo tanto es información más precisa.
  - Si ninguna de las dos es subclase de la otra, se buscan relaciones $R(x_i,y_i)$ y $R'(y_i,x_i)$ tales que los individuales x_i pertenezcan a clase1 y los y_i a clase2. Si existen R o R', se retornan los nombres de los individuales x_i, porque en general, los términos de las búsquedas que están más a la izquierda (en este caso clase1), son más importantes. En otro caso se retornan los nombres de las dos clases, ya que no se pudo obtener información más precisa. Por ejemplo, si las clases son *Ciudad* y *Provincia*, y suponiendo que existe una relación *parte-de* entre individuales de ellas, entonces se retornan los nombres de las ciudades que son parte de una provincia.
- Si los términos identificados son una clase y un individual.
  - Si el individual identificado es instancia de la clase identificada, retorna el nombre del individual, ya que la clase en este caso sería información redundante. Por ejemplo, si la clase es *País* y el individual es *Argentina* (instancia de *País*), se retorna *Argentina*, ya que puede inferirse que *Argentina* es un país.
  - Si el individual no pertenece a la clase identificada, se buscan las relaciones $R(x_i,individual)$ y las $R'(individual,x_i)$ tales que los individuales x_i pertenezcan a la clase identificada. En caso de existir tales relaciones, se retorna los nombres de los x_i, en otro caso retorna *null*. Por ejemplo, si la clase es *Ciudad*, el individual es *Argentina* y R es la relación *parte-de*, entonces se retornarían los nombres de las ciudades que son parte de Argentina.
- Si los términos identificados son dos individuales (individual1 e individual2):
  - Si existe una relación entre ambos individuales, retorna individual1, porque se decide considerar que el primer término de la búsqueda es el más importante; en otro caso se retorna *null*.

En la Sección 6.3. se presenta un ejemplo de la expansión de la consulta luego de la descripción del prototipo.

6.2. Prototipo

Para la implementación del prototipo se optó por utilizar el motor de búsqueda Nutch[26]. Este buscador está basado en la librería Apache Lucene de recuperación de información y permite extender su funcionalidad mediante plugins. Al ser software libre se puede disponer del código fuente sin restricciones y al estar desarrollado en Java se gana portabilidad a los sistemas operativos más conocidos.

Para la *interacción con la ontología* se utilizó la API para ontologías del framework Jena2[27], en conjunción con el lenguaje de consultas SPARQL[28]. Para realizar los razonamientos se utilizó la API para inferencias y los razonadores de Jena2. El *stemming en español* fue implementado a partir de los proyectos Apache Lucene y Snowball[29]. El algoritmo de stemming utilizado es una adaptación del algoritmo de Porter, y se utiliza tanto en el indexado como en el tratamiento de las consultas [Porter, 2001].

El *procesamiento de los metadatos,* la *indexación con vocabulario controlado*, el procedimiento de *expansión de consultas* y el *stemming en español* se implementaron como plugins de Nutch.

Además, a la *interfaz gráfica* del mismo, se agregaron dos funcionalidades para facilitarle al usuario la tarea de búsqueda, como se muestra en la Figura 6.4. Una de ellas es el campo *búsquedas relacionadas*, donde se da la opción al usuario de hacer nuevas búsquedas, a partir de información inferida de la ontología para la consulta ingresada. Para la desambiguación en el refinamiento semántico, se optó por agregar en la interfaz la funcionalidad *otras búsquedas*, que trata los términos ambiguos en las consultas del usuario. Los problemas de ambigüedad aparecen cuando se trata de determinar con qué recurso de la ontología asociar una palabra, cuando hay más de un candidato para hacerlo. Para ello, se presenta al usuario una serie de listas desplegables con las cuales puede reformular la consulta y realizar una búsqueda más precisa. Por ejemplo, si el usuario introduce *Bariloche*, en este menú desplegable se muestran las opciones *San Carlos de Bariloche* y *Asociación*

[26] Nutch, Open Source Search: http://lucene.apache.org/nutch/
[27] Jena2, A Semantic Web Framework for Java: http://jena.sourceforge.net/
[28] SPARQL Query Language for RDF: http://www.w3.org/TR/rdf-sparql-query/
[29] Snowball, a language for stemming algorithms: http://snowball.tartarus.org/

Paleontológica Bariloche, la primera se refiere a la ciudad y la segunda a un museo.

Figura 6.4. Interfaz de usuario del prototipo

El prototipo descrito aquí fue implementado por el alumno Adrián Ponce dentro del marco del Proyecto de Investigación y Desarrollo: “Búsqueda en bases de datos de texto” [Deco, 2007-2010], bajo mi dirección.

6.3. Ejemplo

Supongamos que el usuario ingresa la consulta “*localidad de la provincia de misiones*”. Para resolverla, se ignoran las palabras no significativas o stopwords (“de” y “la”) de la consulta y se mantienen los términos restantes: “localidad”, “provincia” y “misiones”. Luego, se llevan estos términos a su forma raíz (stemming), resultando “local”, “provinci” y “mision”. Suponiendo que *ciudad* es el término del vocabulario controlado asociado a *localidad*, se agrega su forma raíz “ciud” a la expresión de búsqueda. Se agregan términos adicionales, generados a partir de la identificación de clases o individuales de la ontología. En este ejemplo, se identifican las clases *Ciudad*, *Provincia* y el individual correspondiente a la

provincia de Misiones, *pmisiones*. Además, se agregan las ciudades *eldorado*, *obera*, *posadas* y *puertoiguazu*, obtenidas por inferencia a partir de la ontología.

Se consulta al índice en los campos *url*, *anchor*, *title* y *host* por la ocurrencia de los términos “local”, “provinci” y “mision”. En el campo *content*, que es el que contiene los términos en vocabulario controlado, se consulta la ocurrencia de “ciud”, “provinci” y “mision”. Cabe hacer notar que en la consulta al campo *content* se reemplaza “local” por “ciud” que es el término asociado en el vocabulario controlado. En el campo *category*, se consultan las ocurrencias de “ciudad”, “provincia”, “pmisiones”, “eldorado”, “obera”, “posadas” y “puertoiguazu” que son los términos obtenidos de la ontología.

En la Figura 6.5, se muestra la consulta enviada al prototipo. Aquí, el símbolo “+” antepuesto significa que son cláusulas requeridas. Es decir, un documento para ser parte de la respuesta, debe cumplir las tres primeras cláusulas y puede no cumplir la última.

```
+(url:local anchor:local content:ciud title:local host:local category:ciudad category:provincia category:pmisiones)
+(url:provinci anchor:provinci content:provinci title:provinci host:provinci category:ciudad category:provincia category:pmisiones)
+(url:mision anchor:mision content:mision title:mision host:mision category:ciudad category:provincia category:pmisiones)
category:eldorado category:obera category:posadas category:puertoiguazu
```

Figura 6.5: Consulta final de ejemplo

La utilización del campo *content* y su inclusión en el índice, mejora la precisión de los resultados dado que contiene términos del vocabulario controlado definidos para el dominio Turismo. La consulta realizada en el resto de campos utilizando términos libres y el agregado de términos obtenidos por inferencia a partir de la ontología, aumentan la cantidad de documentos recuperados.

6.4. Resultados de la experimentación

Para la evaluación se utilizó el sitio web relativo al turismo en la República

Argentina a partir del cual se construyó la ontología. Se preparó un corpus anotado de 2500 páginas web. Se realizaron 10 consultas, calculando la precisión sobre los primeros 10, 30 y 50 resultados, y se determinó la cantidad de documentos recuperados en cada consulta. Estas consultas se resolvieron con distintas configuraciones, a los efectos de observar cómo reacciona el sistema con el agregado de cierta funcionalidad. Estas configuraciones son: el buscador con su configuración por defecto (Nutch); el buscador con el agregado del vocabulario controlado (Nutch VC); con la utilización de un algoritmo de stemming para el español (Nutch+Stem); con la utilización simultánea de vocabulario controlado y stemming de palabras clave (Nutch+VC+Stem); y por último, cuando se utiliza el refinamiento semántico para la expansión de las consultas (Nutch+QE). La Tabla 6.1 muestra la precisión y la cantidad de documentos recuperados en cada caso.

Se observa que en Nutch VC, en general, hay muy poco aumento de la cantidad de documentos recuperados respecto a la configuración por defecto del buscador (Nutch). Esto indica que el vocabulario utilizado en los documentos es bastante consistente respecto a los términos de las consultas ingresadas. En cuanto a la precisión, tampoco se observan cambios notables entre ambas configuraciones.

La incorporación del algoritmo de stemming (Nutch+Stem) logra un aumento considerable en la cantidad de documentos recuperados, aunque se observa una ligera pérdida en la precisión para ciertas búsquedas, y una pequeña ganancia para otras.

La combinación del uso de vocabulario controlado y de stemming (Nutch+VC+Stem) logra aumentar la cantidad de documentos recuperados respecto a la configuración por defecto (Nutch). En cuanto a la precisión, se pueden hacer, en este caso, observaciones similares a Nutch+Stem, en la que no puede establecerse una tendencia respecto a la precisión de las búsquedas para las consultas realizadas.

Por último, la configuración Nutch+QE, que incorpora el refinamiento semántico, aumenta aún más la cantidad de documentos recuperados al incluir aquellas páginas cuyo contenido traten acerca de los conceptos identificados en la consulta del usuario, y que por lo tanto pueden ser páginas que no incluyan directamente los términos ingresados por el usuario. También puede apreciarse un aumento significativo en la precisión de las búsquedas para 10, 30 y 50 resultados, logrando obtener para algunos casos el 100% de precisión.

Consulta	Nutch				Nutch VC				Nutch + Stem				Nutch + VC + Stem				Nutch + QE			
	Precision			Recall*	Precision			Recall*	Precision			Recall*	Precision			Recall*	Precision			Recall*
	10	30	50		10	30	50		10	30	50		10	30	50		10	30	50	
gaucho	0.40	0.47	0.28	30	0.40	0.47	0.28	30	0.80	0.60	0.62	51	0.80	0.60	0.62	51	0.80	0.60	0.62	51
deporte	0.80	0.80	0.80	68	0.80	0.80	0.80	68	1.00	0.63	0.54	1164	1.00	0.63	0.54	1164	1.00	1.00	1.00	1171
transporte	0.70	0.30	0.20	34	0.70	0.47	0.30	70	0.30	0.27	0.18	81	0.50	0.43	0.30	77	1.00	0.80	0.56	92
ciudad argentina	0.50	0.53	0.34	1118	0.40	0.57	0.34	1224	0.30	0.47	0.34	1078	0.30	0.47	0.36	1185	1.00	1.00	1.00	1197
turismo "san carlos de bariloche"	0.60	0.50	0.36	41	0.60	0.50	0.36	41	0.80	0.53	0.36	41	0.80	0.53	0.36	41	1.00	0.67	0.58	477
museo litoral	0.50	0.23	0.14	50	0.50	0.23	0.14	50	0.20	0.27	0.16	117	0.20	0.27	0.16	117	0.80	0.27	0.16	172
turismo aventura argentina	0.80	0.43	0.30	1047	0.80	0.43	0.30	1047	0.80	0.43	0.34	1046	0.80	0.43	0.34	1046	1.00	1.00	0.92	1379
atractivo natural argentina	0.40	0.40	0.30	34	0.40	0.40	0.30	34	0.30	0.27	0.26	106	0.30	0.23	0.26	107	1.00	1.00	1.00	155
región turística argentina	0.90	0.30	0.18	190	0.90	0.30	0.18	219	0.70	0.30	0.18	281	0.60	0.30	0.18	372	1.00	0.33	0.20	424
excursión región litoral argentina	0.20	0.07	0.04	3	0.40	0.13	0.08	10	0.00	0.07	0.04	23	0.00	0.13	0.08	30	0.40	0.53	0.34	218
Promedio	**0.58**	**0.40**	**0.29**	**261.50**	**0.59**	**0.43**	**0.31**	**279.30**	**0.52**	**0.38**	**0.30**	**398.80**	**0.53**	**0.40**	**0.32**	**419.00**	**0.90**	**0.72**	**0.64**	**533.60**

Tabla 6.1: Resultados de la experimentación

Cuando el usuario consulta por términos que no están en la ontología, como es el caso de la consulta “gaucho”, se obtienen los mismos valores de precisión y de cantidad de documentos recuperados con refinamiento que con la configuración Nutch+VC+Stem. Esto se debe a que al no poder asociar el término “gaucho” con ningún concepto/individual de la ontología, durante la expansión de consultas no se agregan términos del tipo *category* que aumenten la cantidad de documentos recuperados o que mejoren la precisión.

La Figura 6.6 muestra el promedio de la cantidad de documentos recuperados para cada una de las configuraciones. Para la configuración que utiliza el refinamiento semántico para la expansión de consultas (Nutch+QE), se observa que la cantidad de documentos recuperados casi duplica la cantidad obtenida con la configuración por defecto (Nutch).

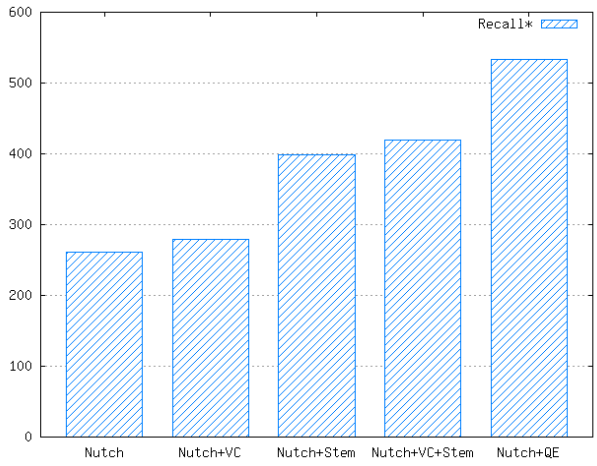


Figura 6.6. Promedio de la Cantidad de documentos recuperados con las distintas configuraciones

En la Figura 6.7 se observan los promedios de los valores de precisión para cada una de las configuraciones para los primeros 10, 30 y 50 resultados. En los tres casos, la configuración con refinamiento semántico (Nutch+QE) presenta un incremento cercano al 33% en la precisión de las búsquedas respecto a la configuración por defecto (Nutch).

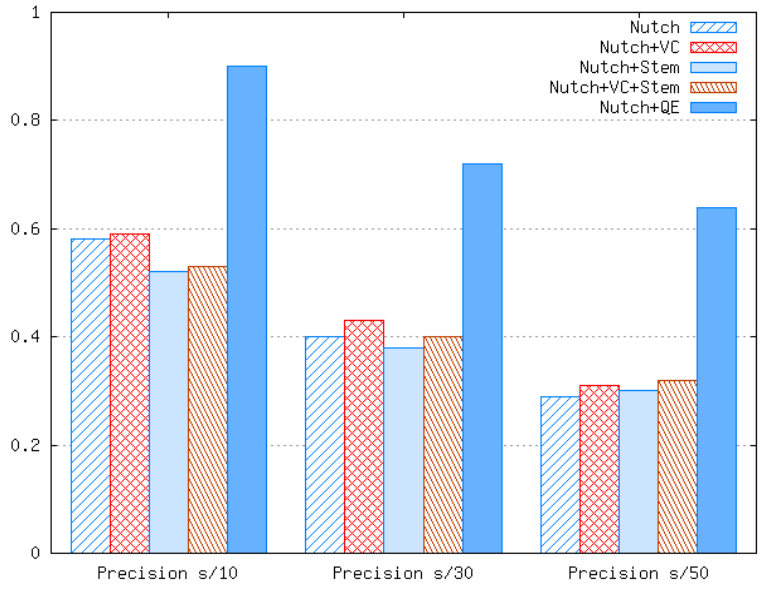


Figura 6.7. Promedio de la Precisión en los primeros 10, 30 y 50 resultados

Por lo tanto, la incorporación del refinamiento semántico al buscador presenta mejores valores de precisión y de la cantidad de documentos recuperados respecto al uso del buscador con su configuración por defecto. Esta propuesta y los resultados de la experiencia fueron publicados en los proceedings del Congreso Argentino de Ciencias de la Computación [Ponce et al., 2008].

Además, la velocidad en la recuperación y la calidad de los resultados son dos propiedades necesarias en cualquier sistema de búsqueda. La incorporación del refinamiento semántico en un motor de búsqueda de un sitio web mejora la calidad de los resultados. La velocidad en la respuesta se logra con una mejora en el índice que consulta el buscador. En [Deco et al., 2007b] se presenta el índice XM-Tree que utiliza las cualidades de los índices sobre espacios métricos. Esto permite la recuperación efectiva y eficiente de objetos. Efectiva, porque los resultados tienen un alto grado de exactitud por propiedades del espacio y del índice; y eficiente porque los índices se construyen para reducir el número de cálculos y objetos revisados. El XM-tree permite la inserción dinámica de nuevos datos, reduce los costos de búsqueda con distancias precalculadas y podas, y utiliza una cantidad de espacio tolerable, lo que lo hace apto para el extenso y dinámico entorno Web. La propuesta emplea la norma L_2 como distancia de indexado y resuelve las búsquedas aplicando como criterio de similitud la norma L_∞.

Entonces, la utilización del índice XM-Tree en el motor de búsqueda, en conjunto con el refinamiento semántico, brinda tanto velocidad como calidad en los resultados de una búsqueda.

Este trabajo fue seleccionado como uno de los mejores del Congreso Argentino de Ciencias de la Computación, y fue invitado para ser publicado en una nueva versión, en el Journal of Computer Science & Technology [Deco et al., 2008d].

6.5. Conclusiones

Los resultados obtenidos durante la experimentación muestran una mejora en la efectividad de las búsquedas al implementar el refinamiento semántico en el buscador del sitio web, con respecto al uso del buscador en su forma estándar. En la experimentación realizada se obtuvo un incremento cercano al 33% en la Precisión y un valor de la cantidad de documentos recuperados aproximadamente igual al doble.

La utilización del vocabulario controlado ayuda a mantener la consistencia del vocabulario utilizado en las páginas del sitio web. Además, reduce el problema de la sinonimia de términos.

El agregado de los metadatos en las páginas del sitio, permite hacer una clasificación de los documentos. Este enriquecimiento semántico posibilita la recuperación de páginas que podrían ser consideradas no relevantes. Una alternativa, siguiendo la idea propuesta por [Nagypàl, 2005], es generar los metadatos en forma automática a partir de la ontología durante el indexado, con lo cual se evita su inserción en los documentos.

La ontología se utiliza para definir el vocabulario controlado, caracterizar el contenido del sitio por medio de los metadatos, para la expansión de consultas y para tratar la polisemia de términos.

Queda abierto evaluar la reutilización de ontologías existentes para describir el dominio de aplicación, así como la utilización de bases de datos léxicas existentes como tabla de términos equivalentes asociada al vocabulario controlado.

Capítulo 7: Utilización del Refinamiento Semántico en un Sistema Recomendador para la Búsqueda de Recursos Educativos

7.1. Introducción

En el dominio de la educación existe gran cantidad y diversidad de material multimedia que puede contribuir al proceso enseñanza-aprendizaje. En particular, con el desarrollo de la Web y su utilización masiva, se tiene una amplia gama de posibilidades de acceso a material útil e interesante para ser empleado tanto por un alumno que desea aprender un tema, o por un docente que desea preparar material didáctico. Sin embargo, se advierte una sobrecarga de información que obliga a estos usuarios a explorar espacios excesivamente densos, convirtiendo la selección de la información que les interesa en una tarea tediosa, que insume mucho tiempo y que es difícil de realizar sin la asistencia de herramientas de búsqueda intuitivas y eficientes. Sin embargo, un material dado no es el adecuado para todos los usuarios. Esto se debe a que los usuarios poseen distintos estilos de aprendizaje, así como características y preferencias personales, que deberían ser consideradas en el momento de la búsqueda.

En los últimos años, los Sistemas Recomendadores ([Resnick et al., 1997], [Terveen et al., 2001]) surgen para ayudar a resolver este tipo de problema puesto que son capaces de seleccionar, de forma automática y personalizada, el material que mejor se adapte a las preferencias o necesidades de un usuario. Estos sistemas utilizan distintas técnicas para razonar sobre las preferencias de los usuarios (modeladas en perfiles personales) y sobre las descripciones semánticas del material disponible.

Actualmente, se conoce como Objetos de Aprendizaje a todo recurso digital que apoya a la educación y que puede ser reutilizado [Wiley, 2002]. El concepto de Objeto de Aprendizaje (Learning Object) abarca principalmente a un conjunto de materiales digitales los que como unidad o agrupación permiten o facilitan alcanzar un objetivo educacional. Ejemplos de los recursos digitales más pequeños incluyen a imágenes o fotos, cortos de video o audio, pequeñas porciones de texto, ecuaciones, definiciones, animaciones, pequeñas aplicaciones web, entre otros. Ejemplos de

recursos digitales de mayor tamaño son páginas web completas que combinen texto, imágenes y otros medios de comunicación.

En este capítulo se presenta un sistema recomendador en cuya arquitectura general se incluye el refinamiento semántico. El objetivo del sistema recomendador es apoyar a los usuarios a encontrar recursos educativos de acuerdo a sus características y preferencias. El refinamiento semántico se utiliza para construir la estrategia para la búsqueda temática y el sistema recomendador brinda al usuario estos resultados ordenados de acuerdo a su perfil.

La propuesta de utilizar el Refinamiento Semántico como una parte importante en la búsqueda de cursos surge dentro del marco del proyecto “EduCa: Red de Educación con Calidad Cultural” [EDUCA]. Este fue un proyecto conjunto entre grupos de investigación de universidades de Uruguay, Argentina y Brasil que fue evaluado y aprobado por el Fondo Regional para la Innovación Digital en América Latina y el Caribe (FRIDA) y fue uno de los 12 seleccionados sobre 122 proyectos presentados.

El proyecto JARDIN (Just an Assistant foR instructional DesIgN) [JARDIN] se presenta como una continuación del proyecto EduCa. JARDIN es un proyecto conjunto entre grupos de investigación latinoamericanos que tiene por objetivo el desarrollo de una herramienta que facilite la creación, descripción, búsqueda y re-uso de Objetos de Aprendizaje [Motz et al, 2008]. En particular, se plantea el uso de un sistema recomendador como parte de la herramienta para mejorar la recuperación de objetos de aprendizaje. JARDIN fue uno de los cinco proyectos aceptados por la Federación Latinoamericana y del Caribe para la Investigación Colaborativa en Tecnologías de Información y Comunicación: LACCIR (Latin American and Caribbean Collaborative ICT Research)).

7.2. Arquitectura para la Recuperación de Recursos Educativos

Cuando un usuario busca material en distintos repositorios y en la web (por ejemplo: cursos, tutoriales, simulaciones, etc.) espera encontrar lo que busca y que mejor se ajuste a sus preferencias y características. Dada una consulta temática

resuelta por el refinamiento semántico, todos los usuarios reciben la misma respuesta, a pesar de que tengan distintas características personales (como ser material en su idioma materno) y preferencias (como ser recuperar material con más contenido práctico que teórico). Los sistemas recomendadores pueden ayudar a los profesores y a los estudiantes a encontrar los mejores recursos educativos que se ajusten a su perfil, considerando la similitud entre el usuario y las características de los cursos.

Según [Zaiane, 2002], los sistemas recomendadores han surgido con el comercio electrónico pero no se habían aplicado en el dominio educación y propone usarlos en este campo del conocimiento. Siguiendo esta propuesta, posteriormente [Romero et al., 2007] y [Soonthornphisaj et al., 2006] presentan trabajos para el uso de sistemas de recomendación en este dominio. [Romero et al., 2007] plantea el uso de técnicas de minería de datos para recomendar la navegación entre links. [Soonthornphisaj et al., 2006] propone un sistema que integra el material recomendado antes de dárselo al usuario.

La personalización de los resultados se sustenta en los metadatos, tanto del usuario como de los documentos. Los metadatos descriptivos siguen el estándar LOM (Learning Object Metadata) [LOM] y una extensión de este estándar para la consideración de característica culturales, propuesta en [Motz et al., 2005]. Los metadatos permiten evaluar la cercanía de un documento al perfíl del usuario.

Una arquitectura para la recuperación de recursos educativos que le brinde al usuario el material que responda a su necesidad temática y a sus preferencias se presenta en la Figura 7.1. La propuesta de esta arquitectura y sucesivas extensiones se presentan en [Bender et al, 2006], [Casali et al, 2006] y en [Deco et al., 2008e].

La arquitectura propuesta está compuesta por: el Refinador Semántico (RS), un Manejador de Perfiles de Usuario (MPU), y el Recomendador de Recursos Educativos (RRE). Se asume además que existe un Repositorio de Objetos de Aprendizaje (ROA) con recursos educativos enriquecidos con metadatos que describen las características del objeto como ser el tema que cubre, su idioma, cantidad de imágenes, etc.

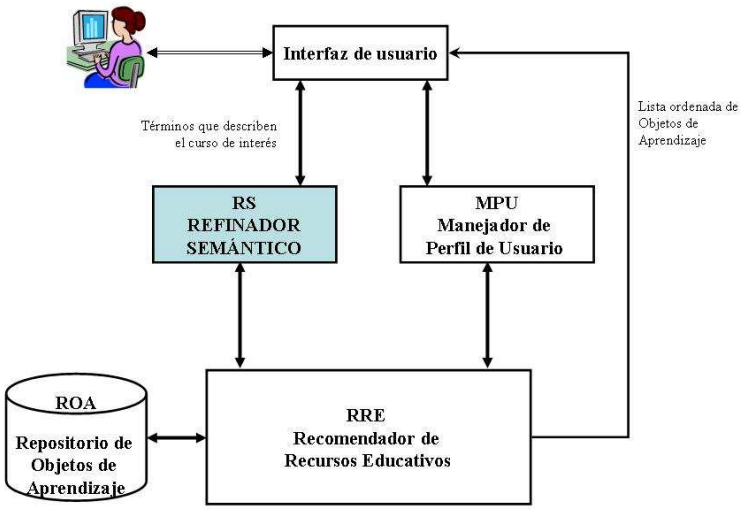


Figura 7.1: Arquitectura del Sistema Recomendador

El Refinador Semántico, propuesto en esta tesis, produce la estrategia de búsqueda asociada al interés del usuario y provee al RRE el conjunto de documentos que satisfacen la búsqueda temática junto con sus metadatos.

El manejador de perfiles de usuario, obtiene y administra los datos personales del usuario, por ejemplo por medio de un conjunto de preguntas dirigidas por una ontología, como se propone en [Motz et al., 2005]. Este módulo provee al RRE con los metadatos correspondientes a las características y preferencias del usuario.

El Recomendador de Recursos Educativos (RRE) evalúa la similitud entre las características de los documentos que satisfacen la búsqueda temática y el perfil del usuario. Como resultado le devuelve al usuario una lista ordenada donde el primer elemento es el más cercano a su perfil.

Se propone el diseño del RRE como un agente BDI graduado [Casali et al, 2005]. El modelo BDI (Beliefs-Desires-Intentions[30]) se elige porque este agente debe decidir cuáles son los mejores recursos a ofrecer al usuario (intenciones) según las características de los recursos educativos (creencias) y las preferencias y

[30] Creencias-Deseos-Intenciones

restricciones del usuario (deseos).

El uso de un modelo intencional como lo es el BDI, permite especificar una arquitectura donde todas las actitudes mentales y sus interacciones pueden ser representadas y pesadas para realizar decisiones más flexibles. Por otro lado, un modelo graduado es adecuado cuando no hay certeza o hay incertidumbre en la forma en que un recurso satisfaría a un usuario y además cuando las preferencias y restricciones del usuario son graduadas.

La representación de los metadatos se propuso en [Bender et al., 2008]. Las características de cada recurso educativo se representan con un vector MD_{Cj} de *n* componentes, donde *n* es la cantidad total de características y cada componente es una tripleta que contiene el nombre de la característica, su valor y su peso. El peso indica la importancia que posee esa característica para que el recurso sea aprovechado por el usuario. Para el desarrollo de un prototipo se consideran las siguientes características: *Difficulty Level, Language, Amount of Practice, Theory and Images, Interactivity* y *Student Participation.* Del mismo modo, se representa el perfil del usuario con un vector MD_U. Los atributos considerados en este caso son: *Knowledge Level, Language (reading, listening, speaking, writing), Mother Language, Attitude,* y *Learning Style.* En la Sección 7.3 se muestra un ejemplo de estos vectores.

Dos problemas que se analizan en [Bender et al., 2008] son el tratamiento de características con valores difusos, y el tratamiento de la ausencia de una o más características. Otro problema que se estudia es que los valores de las características en los vectores MD_U y MD_{Ci} son de distinto tipo: porcentaje (como en el caso de cantidad de imágenes), palabras obtenidas a partir de una lista (como en el caso de Actitud que puede ser Activo, Pasivo, Reactivo), multivaluados (por ejemplo a un usuario le puede interesar más de un idioma), etc. Cada tipo es tratado en una forma diferente. Además, para evaluar la similitud entre los metadatos de un recurso y los del usuario, los conjuntos a comparar deben ser equivalentes. Para esto de define una tabla de mapeo entre ambos. Los vectores se representan como puntos en el espacio n-dimensional y la similitud entre ellos se evalúa calculando la distancia Euclídea. La búsqueda por similitud permite recuperar no sólo aquellos recursos cuyos metadatos coincidan exactamente con los del usuario, sino también aquellos que tengan alguna

semejanza.

Entonces, volviendo a la arquitectura de la Figura 7.1, el RRE calcula todas las distancias de los recursos recuperados de la búsqueda temática y presenta una lista ordenada, donde el primer elemento es el más próximo (es decir, el que tiene menor distancia) a sus preferencias y características. Por lo tanto, se le presentan al usuario todos los recursos que satisfacen su búsqueda temática que es resuelta por el refinamiento semántico y no se descartan elementos. Un recurso educativo que sea muy disímil al perfil del usuario estará el final de la lista, pero no será descartado.

7.3. Ejemplo

Supongamos que María es una estudiante de Ingeniería Industrial que está buscando cursos de Cinemática y que éste es el primer curso de Física que ella toma. Es Argentina, por lo que su lengua materna es el español. Pero, a pesar de tener un nivel bajo de conocimiento del idioma inglés, ella prefiere cursos en este idioma para mejorarlo. Además, es una estudiante pasiva y su estilo de aprendizaje es analítico verbal.

Entonces, su vector de usuario es:

$MD_{María}$ = { (KnowledgeLevel, “Low”, 1),
(English,“Low”,1),
(MotherLanguage,“Spanish”,0.4),
(Attitude,“Passive”,0.1),
(Learning Style,“Analytic Verbal”,0.9) }

Como este es el primer curso de física que realizará, su conocimiento sobre el tema (indicado en la componente *KnowledgeLevel* del vector $MD_{María}$) es bajo; y por esto es muy importante (Peso = 1) que el nivel del curso sea inicial. Como se mencionó, su conocimiento de inglés (componente *English*) es bajo. Pero como quiere mejorar el manejo de este idioma le interesa que el recurso a recuperar esté en

inglés, por lo que su peso es 1, y el peso de material en español es menor (0.4). Su actitud (*Attitude*) es Pasiva porque prefiere leer material y trabajar sola, pero en esta búsqueda no tiene importancia para ella por lo que su peso es bajo (Weight=0.1). Su estilo de aprendizaje (*LearningStyle*) es Analítico Verbal por lo que prefiere recursos con más teoría que práctica y con poca cantidad de imágenes. Esto si es bastante importante para ella por lo que su peso es cercano a 1.

Cuando realiza la búsqueda, María proporciona como término de entrada la palabra *Física*. El Refinador Semántico le muestra la jerarquía conceptual asociada, donde *Cinemática* es un término específico. Ella selecciona este término que luego es expandido semánticamente por este módulo a fin de agregar sinónimos y términos equivalentes en otros idiomas. La estrategia de búsqueda resultante devuelve un conjunto de cuatro recursos de objetos de aprendizaje del repositorio, con sus metadatos MD_{C1}, MD_{C2}, MD_{C3} y MD_{C4}. Por ejemplo, el primer vector MD_{C1} contiene los metadatos de un curso de Cinemática de Nivel Intermedio de dificultad, en idioma Inglés, donde no se requiere interactividad, con mucha teoría, con un 40 % de práctica y pocas imágenes, y es:

MD_{C1} = { (DifficultyLevel, “Intermediate”, 1),
(Language, “English”, 1),
(PracticeAmount, 40%, 1),
(TheoryAmount, 100%, 1),
(ImagesAmount, 10%, 1),
(Interactivity, “No”, 1),
(StudentParticipation, “Individual”, 1) }

Luego de la conversión a R^n, se tienen los valores presentados en la Tabla 7.1.

Curso	Difficulty Level	English	Spanish	PracticeAmount	TheoryAmount	ImagesAmount	Interactivity	StudenParticipation
C_1	0.50	1	0	0.4	1.0	0.1	0.25	0.16
C_2	0.16	1	0	0.5	0.8	0.3	0.25	0.16
C_3	0.83	0	1	0.5	1.0	0.9	0.25	0.16
C_4	0.50	0	1	0.3	0.9	0.8	0.75	0.16

Tabla 7.1: Cursos para la Búsqueda sobre Cinemática

De igual forma, el manejador de perfil del usuario adquiere el perfil de María MU_{Maria} y lo convierte a R^n, obteniendo

$$MU_{Maria} = (0.016; 0.83;0.40;0.15;0.749;0.15;0.016;0.016)$$

El recomendador RRE, a partir del conocimiento que tiene de los cursos que satisfacen la consulta temática y de sus características (Beliefs) y del perfil de usuario que representa sus deseos (Desires) debe decidir cuál es el mejor curso a recomendar (Intention). En esta propuesta la satisfacción del usuario se evalúa con la idea de que “el usuario estará satisfecho en un mayor grado si el curso es más similar a su perfil”. Para esto, se calcula la distancia Euclídea entre los metadatos del usuario y los de los cursos. La Tabla 7.2 muestra los resultados de los cálculos de estas distancias ordenados en forma creciente de valor de distancia.

Curso	Distancia
C2	0.6601
C1	0.7923
C4	1.5223
C3	1,5959

Tabla 7.2.: Orden recomendado de cursos

Así, el sistema recomienda a María el curso C_2 como el más adecuado y además, le da una lista ordenada de otras alternativas como los cursos C_1, C_4 y C_3 en ese orden.

7.4. Conclusiones

En este capítulo se ha presentado un sistema recomendador de recursos educativos donde el refinamiento semántico es un elemento fundamental para la búsqueda temática.

El enfoque fue lograr la personalización de los resultados en la recuperación utilizando los metadatos del usuario y de cada recurso. El perfil del usuario se incorpora en la recomendación a partir de sus preferencias y la importancia relativa de cada una al momento de elegir un recurso educativo. Esto en conjunto con los metadatos de cada recurso educativo constituye la base para el razonamiento del sistema recomendador.

La ventaja de un recomendador es que permite mejorar la recuperación obtenida del refinamiento semántico porque ordena los resultados de distinta forma según el usuario que haya realizado la consulta.

Uno de los mayores problemas encontrados para la experimentación es la falta de metadatos en los documentos. A pesar de que existen repositorios de recursos educativos como FLOR[31] y OERCommons[32], que prevén la inclusión de metadatos en su diseño, estos metadatos por lo general no se encuentran cargados. Por este motivo, se presentó esta propuesta en el llamado a proyectos de la Secretaría de Estado de Ciencia, Tecnología e Innovación de la Provincia de Santa Fe en el marco del Programa 2 de los Programas de Promoción de las Actividades Científico Tecnológicas y de Innovación. El Proyecto “Sistema de Apoyo al Docente en la Búsqueda y Preparación de material didáctico para la enseñanza de las ciencias en las escuelas santafesinas” fue aprobado para su ejecución en 2009 (Resolución 062 del

[31] Federación Latinoamericana de Repositorios - http://ariadne.cti.espol.edu.ec/FederatedClient
[32] Open Educational Resources - http://www.oercommons.org/

12 de diciembre de 2008). En este Proyecto se plantea la creación de un repositorio con material educativo que contenga los metadatos de interés y la implementación de esta arquitectura.

Capítulo 8: Conclusiones y trabajos futuros

8.1. Conclusiones

Una búsqueda de información es óptima cuando todos los documentos recuperados son relevantes y todos los documentos relevantes son recuperados. El maximizar la cantidad de documentos relevantes obtenidos para una consulta depende de la destreza del usuario para preparar la estrategia de búsqueda. Si bien el usuario no tiene por qué conocer técnicas de recuperación de información, la propuesta general de esta tesis es la de mejorar los resultados de su búsqueda por medio de un “especialista” que implementa estas técnicas.

El refinamiento semántico propuesto consiste en: guiar al usuario para desambiguar los conceptos ingresados por él, permitirle seleccionar conceptos jerárquicamente relacionados a fin de precisar los documentos a recuperar y expandir semánticamente los conceptos a fin de aumentar la cantidad de documentos a recuperar. Los recursos lingüísticos que pueden utilizarse para el refinamiento semántico son tesauros, diccionarios, diccionarios multilingües y ontologías. Qué recurso o recursos se pueden utilizar depende del área del conocimiento. Este refinamiento es semiautomático, pues en ciertas tareas se requiere la participación del usuario: la desambiguación de los conceptos y la selección de conceptos jerárquicamente relacionados. Se propuso un refinamiento semiautomático pues se considera que el esfuerzo inicial que se pretende por parte del usuario en estas dos tareas es recompensado evitándole a posteriori la lectura y el descarte de los documentos que no sean de su interés.

La arquitectura propuesta se presenta en la Figura 4.1. Para evaluar el impacto de utilizar el refinamiento semántico para mejorar los resultados de la búsqueda se implementó un prototipo. Para el desarrollo del prototipo se utilizaron estándares y recomendaciones del grupo W3C así como también lenguajes y recursos libres disponibles en la web.

Para la experimentación del refinamiento semántico se utilizó el recurso lingüístico WordNet para consultas de dominio general y el recurso lingüístico

MeSH, especializado en el área salud, para las consultas en un dominio específico del conocimiento. Las experiencias realizadas se presentan en el Capítulo 5.

Cabe destacar que en los resultados generales de ambas experiencias se observa que el refinamiento semántico mejora la precisión en los primeros 50 documentos en 19,03 % en un dominio general del conocimiento utilizando recursos generales y en un 33,50 % en dominio específico utilizando recursos especializados. Esto muestra que la propuesta de refinamiento semántico presentada mejora la recuperación de información de la web al utilizar recursos lingüísticos para la preparación de la estrategia de búsqueda. Además estos resultados mejoran en un dominio específico del conocimiento si se utiliza un recurso especializado.

Sin embargo, en la experimentación con el recurso del dominio general, Wordnet, en algunas consultas aumenta la cantidad de documentos recuperados, pero decrece la precisión. Las razones que pueden explicar esto son: hay muchas relaciones semánticas que no están en WordNet; hay muchos nombres propios que no están incluidos y hay vocabulario específico que no está incluido en WordNet. Esta última razón es una de las mayores debilidades de WordNet para los propósitos de recuperación de información especializada porque abarca todos los temas y no uno específico. Por lo tanto, en búsquedas en áreas específicas del conocimiento, es necesario analizar en cada área qué recurso lingüístico especializado utilizar.

Otro problema tratado en esta tesis es que actualmente, muchos sitios web manejan un caudal de información considerable. Estos sitios suelen tener un buscador propio, con los mismos problemas que ocurren en la búsqueda en la web. En el Capítulo 6 se presenta otro aporte de esta tesis proponiendo mejorar las prestaciones del motor de búsqueda de un sitio web mediante el refinamiento semántico. Para esto, se adaptó la arquitectura propuesta en la Sección 4.1., para la búsqueda en un sitio web y no en toda la web. Si bien la propuesta es aplicable a cualquier dominio del conocimiento, la experimentación se realizó sobre un sitio de turismo. Los resultados obtenidos durante la experimentación muestran una mejora en la efectividad de las búsquedas al implementar el refinamiento semántico en el buscador del sitio web, con respecto al uso del buscador en su forma estándar. En la experimentación realizada se obtuvo un incremento cercano al 33% en la Precisión y un valor de la cantidad de documentos recuperados aproximadamente igual al doble.

En el Capítulo 7 se ha presentado un sistema recomendador de recursos educativos donde el refinamiento semántico es un elemento fundamental para la búsqueda temática. El enfoque fue lograr la personalización de los resultados en la recuperación utilizando los metadatos del usuario y los metadatos de cada recurso. La ventaja de un recomendador es que permite mejorar la recuperación obtenida del refinamiento semántico porque ordena los resultados de distinta forma según el usuario y el momento en que éste haya realizado la consulta.

Aportes adicionales se presentan en el Apéndice 1 y en el Apéndice 2. En el primero se hace un relevamiento de recursos lingüísticos disponibles en línea y en el segundo se discuten los problemas que aparecen en la búsqueda de información multilingüe, con especial atención a distintos recursos lingüísticos que pueden utilizarse, y los problemas que se presentan en la traducción de la consulta. Las experiencias realizadas tuvieron como objetivo evaluar algunos diccionarios multilingües, disponibles en línea, para las traducciones entre los idiomas español, inglés y francés. De estas experiencias se ha observado: en algunos casos la traducción no es bidireccional, muchos diccionarios no tienen traducciones para conceptos formados por varias palabras ni para los sustantivos propios ni para términos específicos o técnicos, además una palabra puede tener varias traducciones distintas. En el refinamiento semántico propuesto en esta tesis, la desambiguación del término que realiza el usuario en la primera etapa de la preparación de la estrategia fija la acepción de interés del término y esto permite mejorar el proceso de traducción de la consulta.

8.2. Problemas Abiertos

En esta sección se presentan posibles mejoras y extensiones a realizar en el refinamiento semántico para la preparación de la estrategia de búsqueda.

- *Preparación para contingencias*.

El refinador semántico resuelve muchos problemas presentados en la preparación de la estrategia de búsqueda: la desambiguación de términos ambiguos o

no específicos, el correcto uso de la disyunción y de la conjunción, el uso correcto de paréntesis, la inclusión de sinónimos y palabras con distintas formas de escritura, la utilización de términos específicos, el uso correcto de la negación y los errores de tecleo. Sin embargo, otras contingencias que se pueden encontrar en la búsqueda son: cómo aumentar la cantidad si no se recupera información suficiente, y cómo reducir la cantidad si se recuperan demasiados documentos.

Queda abierto el problema de que en caso de obtener como resultado pocos o ningún documento porque se ingresaron demasiados conceptos, definir qué concepto quitar de la estrategia de búsqueda a fin de aumentar la cantidad de documentos recuperados. Con respecto al tema de que el usuario utilice términos demasiado específicos, debería detectarse cuál es el término demasiado específico y definirse una forma de moverse en la jerarquía conceptual para realizar un nivel menos de especificación. Otro problema abierto es analizar la incorporación de operadores de proximidad en la estrategia de búsqueda generada por el refinador semántico. Es decir, operadores que permitan recuperar conceptos que estén en un mismo párrafo, o que estén separados por una cierta cantidad de palabras uno de otro.

- *Selección automática del recurso lingüístico adecuado*

La utilización de un perfil de usuario permitiría la selección automática de los recursos lingüísticos más adecuados para la generación de la estrategia de búsqueda. Por ejemplo, si se detecta que el usuario es un médico, es más adecuado utilizar recursos específicos del área salud, como ser el tesauro Mesh, en lugar de un recurso general, como lo es WordNet. El perfil de usuario se puede armar a partir de una plantilla de datos personales y preferencias que complete el usuario y a partir de logs de estrategias anteriores que satisfacieron la necesidad de información de este usuario. Otra posibilidad es armar perfiles de usuario genéricos a partir solamente de estos logs. Por ejemplo, detectando que todo usuario que pidió “cáncer” y “terapia” se refería al área medicina. En este caso, se podría evitar el paso de desambiguación aprendiendo de estrategias anteriores que, si coexisten estas palabras en una consulta, se refieren al área medicina.

- *Extracción automática de conceptos para la estrategia de búsqueda*

En la propuesta presentada en esta tesis, los conceptos que representan el interés de búsqueda, son ingresados uno a uno por el usuario en la forma de palabras claves. Otra forma es que el usuario ingrese su consulta en la forma de una frase escrita en lenguaje natural y se extraigan automáticamente los conceptos iniciales para el refinamiento semántico.

Para esto, se debe segmentar y etiquetar el texto mediante un analizador morfológico. Un término puede ser etiquetado con más de una etiqueta morfosintáctica, por ejemplo *trabajo* puede ser un sustantivo o un verbo conjugado. En estos casos, la desambiguación con respecto a la etiqueta se efectúa con información lingüística y se complementa con técnicas estadísticas. Esto permite descartar aquellos términos de la frase ingresada por el usuario que no sean sustantivos, ya que éstos son los que generalmente se utilizan como palabras claves de búsqueda. Además, el análisis morfosintáctico del texto permite detectar términos irrelevantes para la búsqueda, como ser artículos, así como detectar construcciones que tienen significado como unidad y no por separado, como ser el caso de sustantivos compuestos. Una propuesta de diseño de una base de datos para el análisis morfosintáctico de texto se presenta en [Deco et al., 2008] [Deco et al., 2008c].

- *Utilización de ontologías con axiomas*.

En el presente trabajo se utilizaron ontologías sin axiomas, también llamadas ontologías livianas. Otra posibilidad es utilizar ontologías con axiomas, y por lo tanto poder realizar inferencias. Es decir, además de los conceptos jerárquicamente relacionados o sinónimos, incorporar a la estrategia de búsqueda nuevos conceptos obtenidos a través de la inferencia.

- *Utilización de Feedback de Relevancia*.

Una de las tareas que podría incorporase al refinador es realizar un feedback de relevancia en base a documentos identificados por el usuario como relevantes. De esta forma se pueden encontrar palabras en dichos documentos e incorporarlas a la

estrategia de búsqueda. El problema aquí es determinar cómo y qué términos extraer de los documentos señalados como relevantes. Una posibilidad es que el usuario los elija y otra es que esto se haga automáticamente a partir de estadísticas y agrupación de las palabras que aparecen en estos documentos.

- *Enfoque de agentes*.

Como se propuso en el Capítulo 8, el refinador semántico puede estar inmerso dentro de un sistema recomendador. Este tipo de sistemas generalmente se modelan como sistemas multiagentes. Por lo tanto el refinador semántico podría modelarse utilizando el enfoque de agentes.

Los agentes surgen dentro del campo de la Inteligencia Artificial y representan una nueva forma de analizar, diseñar e implementar sistemas de software complejos [Jennings et al. 1998]. Se puede definir un agente como una aplicación informática con capacidad para decidir cómo actuar para alcanzar sus objetivos. Un agente inteligente puede funcionar fiablemente en un entorno rápidamente cambiante e impredecible, como es la web. Pueden configurarse con diferentes perfiles para tomar decisiones de acuerdo a las necesidades del usuario y hacer tareas más específicas y personalizadas.

8.3. Publicaciones realizadas durante el transcurso de la tesis

- Motz, R., Deco, C., Bender, C., Saer, J., Chiari, M. Refinamiento semántico para la recuperación de información desde la web. En Proceedings de Iberamia 2004. IX Ibero-American Workshops on Artificial Intelligence. pp 172-179. ISBN 968-863-786-6. Puebla, México, November 2004.
- Motz, R., Guzmán, J., Deco C. and Bender, C. Applying ontologies to educational resources retrieval driven by cultural aspects. Journal of Computer Science & Technology. ISSN 1666-6038. JCS&T Vol 5, N° 4, pp 279-284, December 2005.
- Deco, C., Bender, C., Saer, J., Chiari, M., Motz, R. Semantic refinement for web

information retrieval. In Proceedings of the 3rd Latin American Web Congress. IEEE Press. pp 106-110, 2005.

- Deco, C., Bender, C., Plüss, J., Dallosta, A., Ramírez, M. Marchionno, P., Pierángeli, G. Una interfaz para mejorar la búsqueda de información en la web mediante la expansión semiautomática de la consulta. Proceedings del Simposio Argentino de Informática y Salud, JAIIO-SIS. ISSN 16661141. 2005
- Deco, C., Bender, C., Saer, J., Chiari, M. Expansión de consultas utilizando recursos lingüísticos para mejorar la recuperación de información en la web. En Víctor M. Castel, Comp. (2005) Desarrollo, implementación y utilización de modelos para el procesamiento automático de textos. Mendoza: Editorial de la Facultad de Filosofía y Letras, UNCuyo: 35-46. ISBN del soporte Internet: 987-575-019-0
- Deco, C., Bender, C., Plüss, J., Dallosta, A., Ramírez, M. Un Sistema para Mejorar la Recuperación de Información Médica en la Web mediante la Expansión Semiautomática de la Consulta. En Revista Española de Informática y Salud. Nro 57, Junio 2006, pp.91-97. Ed. Sociedad Española de Informática de la Salud. España. ISSN 1579-8070. 2006.
- Casali, A., Deco, C., Bender, C. and Motz, R. A multiagent approach to educational resources retrieval. En Proceedings del Workshop on Artificial Intelligence for Education (WAIFE), en el marco del 35º Jornadas Argentinas de Informática e Investigación Operativa. ISSN 1850 2784. pp 35-41. Mendoza, Argentina. Septiembre 2006.
- Bender, C., Deco, C., Casali, A., Motz R. Una plataforma multiagente para la búsqueda de recursos educacionales considerando aspectos culturales. En Revista TE&ET (Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología). Diciembre 2006. Vol. 1 Nro. 1. Editorial Responsable Red de Universidades Nacionales con Carreras de Informática (RedUNCI). ISSN 1850-9959. pp 20-29. 2006.
- Bender C., Deco C., Plüss J., Dallosta A., Ramírez M. Un sistema de búsqueda asistida de información médica en la web. En Proceedings del Simposio Argentino de Informática y Salud en el marco del 35º Jornadas Argentinas de

Informática e Investigación Operativa. ISSN 1850 2822. pp 19-28. Mendoza, Argentina. Septiembre 2006.

- Deco, C., Bender, C., Chiari, M. Problemas de la traducción de la consulta en la búsqueda de información multilingüe. En Revista INFOSUR. Vol 1 Nro 1. Universidad Nacional de Rosario. ISSN 1851 1996. pp 39-50. Junio 2007.
- Deco, C., Pierángeli, G., Bender, C., Reyes, N. XM-Tree: Un nuevo índice para recuperación de información en la web. En Proceedings del IV Workshop de Ingeniería de Software y Bases de Datos en el marco del XIII Congreso Argentino de Ciencias de la Computación, CACIC 2007. pp. 656-667, ISBN 978-950-656-109-3. Corrientes, Argentina, octubre 2007.
- Bender, C., Motz, R., Deco, C., Saer, J. Recuperación personalizada de e-cursos. En Proceedings del IX Congreso Iberoamericano de Informática Educativa, RIBIE 2008. Caracas, Venezuela, marzo 2008.
- Deco C., Bender C., Solana Z. Base de datos para el análisis morfosintáctico de un corpus con anotación lingüística. En Proceedings del XI Congreso de la Sociedad Argentina de Lingüística (SAL), dentro de las Jornadas Argentinas de Lingüística Informática: Modelización e Ingeniería, Santa Fe, Argentina, abril 2008
- Deco, C., Bender, F. Severino Guimpel C. Reyes, N. Recuperación de información en bases de datos de texto. En Proceedings del Workshop de Investigadores en Ciencias de la Computación WICC 2008. General Pico, La Pampa, Argentina. Mayo 2008
- Deco C., Bender C., Solana Z. Base de datos para el análisis morfosintáctico de un corpus con anotación lingüística. En Revista INFOSUR. Año 2 Nro 2. Universidad Nacional de Rosario. ISSN 1851 1996. pp 51-60. Septiembre 2008. Artículo seleccionado de las Jornadas Argentinas de Lingüística Informática: Modelización e Ingeniería, realizadas en el marco del XI Congreso de la Sociedad Argentina de Lingüística (SAL).
- Ponce, A., Deco, C., Bender, C. Proposal of an ontology based web search engine. En Proceedings del Workshop de Bases de Datos en el marco del XIV Congreso Argentino de Ciencias de la Computación, CACIC 2008. Chilecito,

Argentina, octubre 2008.

- Deco, C., Pierángeli, G., Bender, C. Reyes, N. XM-Tree, a new index for Web Information Retrieval. Journal of Computer Science and Technology (JCS&T). Vol. 8, Nro 2, pp 78-84. July 2008.
- Motz, R., Viola, A., Palazzo, J., Valdení, J., Ochoa, X., Deco, C., Casali, A., Bender, C., Pérez M., Brunetto, A., Proença, M. Project JARDIN: Just an assistant for instructional design. En Proceedings de la 3ra. Conferencia Latinoamericana de Objetos de Aprendizaje LACLO 2008. México. Octubre 2008
- Deco, C., Bender, C., Casali, A. Motz, R. Design of a Recommender Educational System. En Proceedings de la 3ra. Conferencia Latinoamericana de Objetos de Aprendizaje LACLO 2008. México. Octubre 2008. Trabajo premiado entre los cinco mejores del Congreso.

Bibliografia

[Allan et al., 2002] Allan, J., Aslam, J., Belkin, N., Buckley, C., et al. Report of the Workshop on Challenges in Information Retrieval and Language Modeling. Center for Intelligent Information Retrieval, Universidad de Massachusetts, Setiembre 2002.

[Baeza, 1998] Baeza-Yates, R. A.. Searching the Web: Challenges and Partial Solutions. Depto. de Ciencias de la Computación. Universidad de Chile. Proyecto VII.13.AMYRI – CYTED. 1998.

[Baeza et al., 1999] Baeza-Yates, R., Ribeiro-Neto, B. (eds.), Modern Information Retrieval. 1999, New York. ACM Press.

[Ballesteros & Croft, 1996] Ballesteros, L. & Croft, W. B. Dictionary Methods for Cross-Lingual Information Retrieval. In Database and Expert Systems Applications, pages 791–801, 1996.

[Ballesteros & Croft, 1997] Ballesteros, L. & Croft, W. B. Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval. In Research and Development in Information Retrieval, pages 84–91, 1997.

[Ballesteros, 2001] Ballesteros, L. Resolving ambiguity for cross-language information retrieval – A dictionary approach. Universidad de Massachusetts, 2001.

[Bear et al., 1998] Bear, J., Israel, D., Petit, J.; Martin, D. Using Information Extraction to Improve Document Retrieval. SRI International, Reporte, Enero, 1998

[Bender et al., 2003] Bender C., Deco C., Motz R. Utilización de ontologías y tesauros para mejorar la recuperación de la información de la web en el área salud. Publicada en el CD de las IX Jornadas Iberoamericanas de Informática, organizadas por la Agencia Española de Cooperación Internacional (AECI), Ministerio de Asuntos Exteriores de España, y la Red Iberoamericana de Tecnologías de Software para la Década del 2000 (RITOS2), Subprograma VII "Electrónica e Informática Aplicadas", Red Temática VII.J, Programa de Ciencia y Tecnología para el Desarrollo (CYTED). Colombia, 11 al 15 de agosto de 2003.

[Bender et al., 2004] Bender, C., Perlo, L., Deco, C., Motz, R. Combining techniques for the classification of web pages resulting from a query. En Proceedings de las XII Jornadas Chilenas de Computación (JCC 2004). III Workshop de Bases de Datos.

Arica, Chile, noviembre de 2004. ISBN 956-7021-18-X.

[Bender et al., 2005] C. Bender, C. Deco y L. Perló. Clasificación de páginas web como posprocesamiento a la recuperación de la información. En Víctor M. Castel, Comp. (2005) Desarrollo, implementación y utilización de modelos para el procesamiento automático de textos. Mendoza: Editorial de la Facultad de Filosofía y Letras, UNCuyo: 1-12. ISBN del soporte Internet: 987-575-019-0

[Bender et al., 2006a] Bender, C., Deco, C., Casali, A., Motz R. Una plataforma multiagente para la búsqueda de recursos educacionales considerando aspectos culturales. En Revista TE&ET (Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología). Diciembre 2006. Vol. 1 Nro. 1. Editorial Responsable Red de Universidades Nacionales con Carreras de Informática (RedUNCI). ISSN 1850-9959. pp 20-29. 2006.

[Bender et al., 2006b] .Bender C., Deco C., Plüss J., Dallosta A., Ramírez M. Un sistema de búsqueda asistida de información médica en la web. En Proceedings del Simposio Argentino de Informática y Salud en el marco del 35º Jornadas Argentinas de Informática e Investigación Operativa. ISSN 1850 2822. pp 19-28. Mendoza, Argentina. Septiembre 2006.

[Bender et al., 2008] Bender, C., Motz, R., Deco, C., Saer, J. Recuperación personalizada de e-cursos. En Proceedings del IX Congreso Iberoamericano de Informática Educativa, RIBIE 2008. Caracas, Venezuela, marzo 2008.

[Berners-Lee, 2001] Berners-Lee T., Hendler J., Lassila O, The Semantic Web. Scientific American, mayo 2001. 284(5): pp 34-43.

[Bireme] Bireme, Biblioteca Virtual en Salud, www.bireme.br/

[Bollacker et al., 1998] Bollacker, K., Lawrence, S., Lee Giles, C. Citeseer: An autonomous web agent for automatic retrieval and identification of interesting publications. In Proceedings of the Second International Conference on Autonomous Agents, pages 116-113, ACM Press. New York, 1998.

[Boughanem et al., 2002] Boughanem, M., Chrisment, C., Nassr, N. Investigation on Disambiguation in CLIR Aligned Corpus and Bi-directional Translation-Based Strategies. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001, volume

2406 of LNCS. Springer. 2002.

[Broekstra et al., 2002] Broekstra, J. Klein, M., Decker, S., van Harmelen, F., Horrocks, I. Enabling knowledge representation on the Web by extending RDF Schema. En Computer Networks 39, pp. 609-634, 2002.

[Carpineto et al., 2002] Carpineto, C., Romano G., Giannini, V., Improving retrieval feedback with multiple term-ranking function combination. TOIS 20(3):259-290, 2002.

[Casali et al, 2005] Casali, A., Godo Ll. and Sierra C. Graded BDI Models For Agent Architectures. J. Leite and P. Torroni (Eds.) CLIMA V, LNAI 3487, 126-143, 2005.

[Casali et al., 2006] Casali, A., Deco, C., Bender, C. and Motz, R. A multiagent approach to educational resources retrieval. En Proceedings del Workshop on Artificial Intelligence for Education (WAIFE), en el marco del 35º Jornadas Argentinas de Informática e Investigación Operativa. ISSN 1850 2784. pp 35-41. Mendoza, Argentina. Septiembre 2006.

[Chen et al., 1998] Chen, L., Sycara, K. WebMate: A Personal Agent for Browsing and Searching. Autonomous Agents. Pages 132--139. ACM Press, 1998.

[CLIR] Cross-language information retrieval project, Universidad de Maryland, College Park: www.clis.umd.edu/dlrg. Página de recursos: www.clis.umd.edu/dlrg/clir/papers.html. Bibliografía sobre el tema: www.clis.umd.edu/dlrg/clir/bibtex.txt.

[Cui et al., 2000] Cui H., Wen J., NIE J., Ma W., Probabilistic Query expansion using Query logs. WWW202, may 7-11, Hawai, USA, ACM 1-58113-449. 2002

[Cunningham, 1999] Cunningham, Hamish. Information Extraction - A User Guide. Research memo CS-99-07. Institute for Language, Speech and Hearing (ILASH), and Department of Computer Science. University of Sheffield, UK. April 1999. http://www.dcs.shef.ac.uk/~hamish

[Davis, 1997] Davis, M. New Experiments in CrossLanguage Text Retrieval at NMSU's Computing Research Lab. In Proceedings of TREC5, pages 447–454. NIST, Gaithesburg, MD, 1997.

[De Rosa et al., 2000] De Rosa, M; Iocchi, L; Nardi, D. Knowledge representation

techniques for information extraction on the Web. Dipartimento di Informatica e Sistemistica, Università di Roma "La Sapienza". http://www.dis.uniroma1.it/%7Eiocchi/pub/webnet98.html.

[Decker et al., 1999] Decker, S., Erdmann, M., Fensel D., Studer. R., Ontobroker: Ontology bases Access to Distributed and Semi-Structured Information. University of Karlssruhe, Institute AIFB. In R. Meersman et al., editor, DS-8: Semantic Issues in Multimedia Systems. Kluwer Academic Publisher, 1999.

[Decker et al., 2000] Decker, S, van Harmelen, F., Brockstra, J., Erdmann, M., Fensel, D. Horrocks, I, Klein, M., Melnik, S. The Semantic Web: on the respective roles of XML and RDF. En IEEE Internet Computing, September/October 2000.

[Deco, 2004-2006] Proyecto RIBS: Recuperación de información basada en semántica. Director del proyecto Claudia Deco. Departamento de Investigación Institucional. Facultad de Química e Ingeniería de Rosario. Universidad Católica Argentina. 2004-2006.

[Deco et al., 2005] Deco, C., Bender, C., Saer, J., Chiari, M., Motz, R. Semantic refinement for web information retrieval. In Proceedings of the 3rd Latin American Web Congress. IEEE Press. pp 106-110, 2005.

[Deco et al., 2005b] C. Deco, C. Bender, J. Plüss, A. Dallosta, M. L. Ramírez, P. Marchionno, G. Pierángeli. Una interfaz para mejorar la búsqueda de información en la web mediante la expansión semiautomática de la consulta. Proceedings del Simposio Argentino de Informática y Salud, JAIIO-SIS. ISSN 16661141. 2005

[Deco et al., 2005c] C. Deco, C. Bender, J. Saer y M. Chiari. Expansión de consultas utilizando recursos lingüísticos para mejorar la recuperación de información en la web. En Víctor M. Castel, Comp. (2005) Desarrollo, implementación y utilización de modelos para el procesamiento automático de textos. Mendoza: Editorial de la Facultad de Filosofía y Letras, UNCuyo: 35-46. ISBN del soporte Internet: 987-575-019-0

[Deco et al., 2006] C. Deco, C. Bender, J. Plüss, A. Dallosta, M. L. Ramírez. Un Sistema para Mejorar la Recuperación de Información Médica en la Web mediante la Expansión Semiautomática de la Consulta. En Revista Española de Informática y Salud. Nro 57, Junio 2006, pp.91-97. Ed. Sociedad Española de Informática de la

Salud. España. ISSN 1579-8070. 2006.

[Deco et al., 2006b] Deco, C., Bender, C., Saer, J., Chiari, M. Recursos lingüísticos en la búsqueda de información multilingüe. En Energeia Cuaderno de Investigación. Año 4 Nro 4. Publicación del Departamento de Investigación Institucional. Facultad de Química e Ingeniería, Universidad Católica Argentina. pp 2-12. ISSN 1668-1622. 2006.

[Deco, 2007-2010] Proyecto de Investigación y Desarrollo (ING201) "Recuperación de Información en Bases de datos de texto". Director del Proyecto: M.Sc. Claudia Deco. Departamento de Sistemas e Informática, Facultad de Ciencias Exactas, Ingeniería y Agrimensura, Universidad Nacional de Rosario. 2007-2010.

[Deco et al., 2007] Deco, C., Bender, C., Chiari, M. Problemas de la traducción de la consulta en la búsqueda de información multilingüe. En Revista INFOSUR. Vol 1 Nro 1. Universidad Nacional de Rosario. ISSN 1851 1996. pp 39-50. Junio 2007.

[Deco et al., 2007b] Deco, C., Pierángeli, G., Bender, C., Reyes, N. XM-Tree: Un nuevo índice para recuperación de información en la web. En Proceedings del IV Workshop de Ingeniería de Software y Bases de Datos en el marco del XIII Congreso Argentino de Ciencias de la Computación, CACIC 2007. pp. 656-667, ISBN 978-950-656-109-3. Corrientes, Argentina, octubre 2007.

[Deco et al., 2008] Deco C., Bender C., Solana Z. Base de datos para el análisis morfosintáctico de un corpus con anotación lingüística. En Proceedings del XI Congreso de la Sociedad Argentina de Lingüística (SAL), dentro de las Jornadas Argentinas de Lingüística Informática: Modelización e Ingeniería, Santa Fe, Argentina, abril 2008

[Deco et al., 2008b] Deco, C., Bender, F. Severino Guimpel C. Reyes, N. Recuperación de información en bases de datos de texto. En Proceedings del Workshop de Investigadores en Ciencias de la Computación WICC 2008. General Pico, La Pampa, Argentina. Mayo 2008

[Deco et al., 2008c] Deco C., Bender C., Solana Z. Base de datos para el análisis morfosintáctico de un corpus con anotación lingüística. En Revista INFOSUR. Año 2 Nro 2. Universidad Nacional de Rosario. ISSN 1851 1996. pp 51-60. Septiembre 2008. Artículo seleccionado de las Jornadas Argentinas de Lingüística Informática:

Modelización e Ingeniería, realizadas en el marco del XI Congreso de la Sociedad Argentina de Lingüística (SAL).

[Deco et al., 2008d] Deco, C., Pierángeli, G., Bender, C. Reyes, N. XM-Tree, a new index for Web Information Retrieval. Journal of Computer Science and Technology (JCS&T). Vol. 8, Nro 2, pp 78-84. July 2008.

[Deco et al., 2008e] Claudia Deco, Cristina Bender, Ana Casali, Regina Motz. Design of a Recommender Educational System. En Proceedings de la 3ra. Conferencia Latinoamericana de Objetos de Aprendizaje LACLO 2008. México. Octubre 2008. Trabajo premiado entre los cinco mejores del Congreso.

[Ding et al., 2004] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs. "Swoogle: a Search and Metadata Engine for the Semantic Web". In Proceedings of the Thirteenth ACM international Conference on information and Knowledge Management, pp. 652-659, New York, USA, 2004.

[Doszkocs, 1982] Doszkocs, T. From research to application: the cite natural language information retrieval system. In Proceedings of the 5th annual ACM conference on Research and development in information retrieval, Springer-Verlag New York, pp 251-262, 1982.

[Dumais et al., 1996] Dumais, S., Landauer, T., M.L.Littman. Automatic Cross-Linguistic information retrieval using latent semantic indexing. In SIGIR'96 Workshop on Cross-Linguistic Information Retrieval, 1996.

[EduCa] EDUCA: Red de Educación con Calidad Cultural. Directora: Dra. Regina Motz, Facultad de Ingeniería de la Universidad de la República, Montevideo, Uruguay. Integrantes: José Palazzo de Oliveira, Universidade Federal do Rio Grande do Sul, Claudia Deco, Universidad Nacional de Rosario, Argentina. Proyecto del Fondo Regional para la Innovación Digital en América Latina y el Caribe (FRIDA). EduCa (Propuesta Nº 72) fue uno de los 12 seleccionados sobre 122 proyectos presentados. Período: 2005-2006.

[Efthimiadis, 1996] Efthimiadis E.N. Query Expansion. In Annual Review of Information Systems and Technology (ARIST), v31, pp 121-187, 1996.

[Eichmann et al., 1998] Eichmann, D., Ruiz, M., Srinivasan, P. Cross-language information retrieval with the UMLS Metathesaurus. Proc. ACM Special Interest

Group on Information Retrieval (SIGIR), ACM Press, NY, 72-80, 1998.

[Eikvil, 1999] Eikvil, L. Information Extraction from World Wide Web. A Survey. Technical Report. Norweign Computing Center, Report Nº 945. July 1999.

[El-Beltagy et al, 2004] S. R. El-Beltagy, A. Rafea, Y. Abdelhamid. "Using Dynamically Acquired Background Knowledge for Information Extraction and Intelligent Search". In Intelligent Agents for Data Mining and Information Retrieval, Idea Group Publishing, pp. 195-206, USA, 2004.

[EuroVoc, 1995] EuroVoc. Thesaurus EuroVoc: Vol 1-3 / European Communities. Luxembourg: Office for Official Publications of the European Communities. 1995.

[Fensel et al., 1998] Fensel, D., Decker, S., Erdmann, M., Studer, R. Ontobroker: The very high idea. In Proceedings of the 11th International Flairs Conference (FLAIRS-98), Sanibal Island, Florida., 1998.

[Fensel et al., 1999] Fensel, D., Angele, J., Decker, S., Erdmann, M., Schnurr, H.-P., Staab, S., Studer, R., Witt, A., On2broker: Semantic-based access to information sources at the WWW. World Conference on the WWW and Internet (WebNet99). Honolulu, Hawaii. 1999. http://citeseer.nj.nec.com/decker99onbroker.html.

[Fensel et al., 2000] D. Fensel, J. Angele, S. Decker, M. Erdmann, H.-P. Schnurr, R. Studer, and A. Witt. Lessons learned from applying AI to the web. International Journal of Cooperative Information Systems, 9(4):361--382, 2000.

[Figuerola et al., 2002] Figuerola, C. G., Gomez, R., Rodrıguez, A. F. Z., Berrocal, J. L. A. Spanish Monolingual Track: The Impact of Stemming on Retrieval. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001, volume 2406 of LNCS, pages 253–261. Springer, 2002.

[Fowler et al., 1999] Fowler, j., Perry, B., Nodine, M., Bargmeyer, B.: Agent-Based Semantic Interoperability. InfoSleuth SIGMOD Record 28:1, pp. 60-67, March, 1999.

[French et al., 2001] French J.C., Powell A.L., Gey F., Perelman N., Exploiting a Controlled Vocabulary to Improve Collection Selection and Retrieval Effectiveness. In Tenth International Conference on Information and Knowledge Management

(CIKM-2001). Atlanta, Georgia, pp 199-206, November 5-10, 2001.

[Gaizauskas et al., 1998]. Gaizauskas, R., Wilks, Y. Information Extraction: Beyond Document Retrieval. Computational Linguistics and Chinese Language Processing, vol. 3, no. 2, pp. 17-60, agosto 1998.

[Gonzalo et al., 1998] Gonzalo, J., Verdejo, F., Chugur, I., Cigarran, J., Indexing with WorNet synsets can improve text retrieval. Proceedings of the COLINA/ACL '98 Workshop on Usage of WordNet for NLP. 1998.

[Grefenstette, 1998] Grefenstette, G. The problem of CrossLanguage Information Retrieval, chapter in Cross-Language Information Retrieval. Kluwer Academic Publishers. 1998.

[Gruber, 1993] Gruber T. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Technical Report KSL-93-04, Knowledge Systems Laboratory, Stanford University, CA, 1993.

[Gruser et al., 1998] Gruser J. R., Raschid L., Vidal M. E., Bright L., Wrapper Generation for Web Accessible Data Sources. Conference on Cooperative Information Systems", pages 14-23,1998.

[Guarino et al., 1999] Guarino, N. et al. OntoSeek: Content-Based Access to the Web. In IEEE Intelligent Systems. Vol 14, Nro. 3, pp. 70-80. Mayo/Junio 1999.

[Guarino et al., 1999] Guarino, N., Masolo, C., Vetere, G., OntoSeek: Content-Based Access to the Web. IEEE Intelligent Systems, 14(3), 70--80, May 1999.

[HONselect] HONselect, HON, Health on the Net Foundation, http://www.hon.ch/HONselect/index_sp.html

[Hull & Grefenstette, 1996] Hull, D. A. & Grefenstette, G. Querying across languages: A dictionary-based approach to multilingual information retrieval. In Proceedings of the 19th International Conference on Research and Development in Information Retrieval, pages 49–57, 1996.

[JARDIN] JARDIN: Just an Assistant foR instructional DesIgN. Proyecto soportado por Latin American and Caribbean Collaborative ICT Research and Microsoft bajo subsidio: LACCIR-RFP2008. Directora: Regina Motz, Universidad de la República, Uruguay. Integrantes: José Palazzo de Oliveira, Universidade Federal do Rio Grande

do Sul, Brazil, Angelica Brunetto, Universidade Estadual de Londrina, Brazil, Claudia Deco, Universidad Nacional de Rosario, Argentina. Xavier Ochoa, Escuela Superior Politécnica del Litoral ESPOL, Ecuador, Miguel Angel Pérez Alvarez, Universidad Nacional Autónoma de México, México.

[Jennings et al., 1998] Jennings, N., Sycara, K., Wooldridge, M., A Roadmap of Agent Research and Development, Autonomous Agents and Multi-Agent Systems. 1, 7-38, Kluwer Academic Publishers, Boston, 1998.

[Kalamboukis, 1995] Kalamboukis, T. Suffix stripping with modern Greek. Program, 29:313–321, 1995.

[Kay, 2001] Kay, M. XSLT - Programmers Reference. 2nd. edition. Wrox Press Ltd. ISBN 1-861005-06-7, 2001.

[Kleinberg, 1998] Kleinberg J., Authoritative Sources in a Hyperlinked Environment. Proc. 9th Ann. ACM-SIAM Symp. Discrete Algorithms, ACM Press, New York, pp.668-677, 1998.

[Kleinberg, 1999] Kleinberg J.M., Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, 46(5):604-632, 1999.

[Kraaij & Pohlmann, 1994] Kraaij, W. & Pohlmann, R. Porter's stemming algorithm for Dutch. In Noordman, L. and de Vroomen, W., editors, Informatiewetenschap, Tilburg, STINFON, 1994.

[Kobayashi et al., 2000] Kobayashi, M., Takeda, K., Information Retrieval on the Web. IBM Research, Tokyo Research Laboratory, IBM, Japan, 2000.

[Lancaster, 1995] F. W. Lancaster. "El Control del Vocabulario en la Recuperación de Información". Ed. Universidad de Valencia, España, 1995.

[LOM] IEEE LOM specification. http://ltsc.ieee.org/wg12.

[López-Ostenero et al., 2003] López-Ostenero, F., Gonzalo, J., Verdejo, F. Búsqueda de información multilingüe: estado del arte. Revista Iberoamericana de Inteligencia Artificial. Nro. 22, pp. 11-35, 2003.

[Losee, 1998] Losee, R., Text Retrieval and Filtering: Analytic Models or Performance, Kluwer, Boston, 1998.

[Lozano Tello, 2001] Lozano Tello, A. Ontologías en la Web semántica.

Departamento de Informática, Universidad de Extremadura, España. I Jornadas de Ingeniería Web ’01.

[Magnini et al., 2000] Magnini, B., Cavaglia, G., Integrating Subject Field Codes into WordNet. Proceedings of LREC-2000, Second International Conference on Language Resources and evaluation, pp. 1413-1418. 2000

[Mandala et al., 1998] Mandala, R., Takenobu T. and Hozumi T., The use of Wordnet in information retrieval. Proceedings of Coling-ACL, 1998.

[Manning et al., 2007] C. Manning, P. Raghavan, H. Schutze. “An Introduction to Information Retrieval”. Cambridge University Press, UK, 2007.

[Martínez et al., 2002] Martínez P., García A. Utilizando recursos lingüísticos para mejora de la recuperación de información en la Web. Revista Iberoamericana de Inteligencia Artificial 16 pp 55-64. 2002.

[Medline] Medline, www.ncbi.nlm.nih.gov/entrez/query.fcgi.

[MeSH] MeSH, Medical Subject Headings, National Library of Medicine, USA, www.ncbi.nlm.nih.gov/entrez/meshbrowser.cgi.

[Meva] Meva, Medical Informatics and Artificial Intelligence, http://www.med-ai.com/meva

[Miller, 1995] Miller, G. A lexical database for English. Communication of the ACM. Vol. 38, Issue 11, pp: 39-41, Nov. 1995.

[Monge et al., 1996] Monge, A., Elkan, C. The webfind tool for finding scientific papers over the Worldwide Web. In Proceedings of the Third International Congress on Computer Science Research, Tijuana, Mexico, 1996.

[Monz & de Rijke, 2001] Monz, C., de Rijke, M. Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German and Italian. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001, volume 2406 of LNCS, pages 262–277. Springer, 2001.

[Motz et al., 2001] Motz R, Do Carmo A., Propuesta para integrar bases de datos que contienen información de la Web. 4to. Workshop Iberoamericano de Ingenieria de Requisitos y Ambientes Software, IDEAS '2001. Costa Rica. 2001.

[Motz et al., 2000] Motz R., Wonsever D., Perelló F. y Ferreiro J. Generación automática de una base de datos con información extraída de la Web. Congreso Argentino de Ciencia de la Computación, Ushuaia, Octubre 2000.

[Motz et al., 2003] Motz R., Deco C., Bender C. Arquitectura de un asistente para la recuperación semántica de referencias bibliográficas en la Web. Anales de la 32 Jornadas Argentinas de Informática e Investigación operativa - JAIIO SIS Simposio Argentino de Informática y Salud. ISSN 1666 1141. Buenos Aires, 2003.

[Motz et al., 2003a] Motz R, Deco C., Bender, C., Manzino C., Perlo L., Ruiz E., von Fürt A. La clasificación en la carga de Web Data Warehouses. Workshop Chileno de Bases de Datos, Jornadas Chilenas de Computación. Chillán, Chile, 2003.

[Motz et al., 2004] Motz, R., Deco, C., Bender, C., Saer, J., Chiari, M. Refinamiento semántico para la recuperación de información desde la web. En Proceedings de Iberamia 2004. IX Ibero-American Workshops on Artificial Intelligence. pp 172-179. ISBN 968-863-786-6. Puebla, México, November 2004.

[Motz et al., 2005] R. Motz, J. Guzmán, C. Deco and C. Bender. Applying ontologies to educational resources retrieval driven by cultural aspects. Journal of Computer Science & Technology. ISSN 1666-6038. JCS&T Vol 5, N° 4, pp 279-284, December 2005.

[Motz et al, 2008] Project JARDIN: Just an assistant for instructional design. R. Motz, A. Viola, J. Palazzo, J. Valdení, X. Ochoa, C. Deco, A. Casali, C. Bender, M. A. Pérez Alvarez, A. Brunetto, M. Proença. En Proceedings de la 3ra. Conferencia Latinoamericana de Objetos de Aprendizaje LACLO 2008. México. Octubre 2008

[Nagypàl, 2005] G. Nagypàl. "Improving Information Retrieval Efectiveness by Using Domain Knowledge Stored in Ontologies". FZI Research Center for Information Technologies at the University of Karlsruhe, pp. 780-789, Germany, 2005.

[National Library of Medicine, 1997] National Library of Medicine. Unified Medical Language System (UMLS). Knowledge Sources, 6th experimental edition, 1997.

[Navigli et al., 2002] Navigli, R., Velardi, P., Automatic Adaptation of WordNet to Domains. 3rd International Language Resources and Evaluation Conference LREC 2002 and ONTOLEX2002 Workshop. Las Palmas, Canary Islands, Spain, May 27th,

2002.

[Navigli et al., 2003] Navigli, R., Velardi, P., An analysis of ontology-based query expansion strategies. Workshop on Adaptive Text Extraction and Mining (ATEM 2003) in the 14th European Conference on Machine Learning (ECML 2003), Cavtat-Dubrovnik, Croatia, September 22-26th, 2003

[Nodine et al., 2000] Nodine, M., Fowler, J., Ksiezyk, T., Perry,B., Taylor M., Unruh, A. Active information gathering in InfoSleuth. In International Journal of Cooperative Information Systems 9:1/2, pp. 3-28. 2000.

[Oard, 1998] Oard, D. W. A comparative study of query and document translation for cross-language information retrieval. In Proceedings of the Third Conference of the Association for Machine Translation in the Americas, 1998.

[OWL] OWL Web Ontology Language 1.0 Reference, http://www.w3.org/TR/2002/WD-owl-ref-20020729/ Consultado el 08-09-05.

[Page et al., 1998] Page L., Brin S.. The PageRank Citation Ranking: Bringing Order to The Web, Standford Digital Library Technologies, Working Paper 1999-0120, Stanford Univ., Palo Alto, Calif., 1998.

[Pirkola, 1998] Pirkola, A. The Effects of Query Structure and Dictionary Setups in Dictionary Based Cross-Language Information Retrieval. In Proceedings of SIGIR'98, pages 55–63, 1998.

[Plüss et al., 2003] Pluss, J., Del Pozo, F.; Hernando, M.E.; Rodriguez, S.; Gómez, E., De Toledo, P. (Universidad Politécnica de Madrid, España); Hernández, C. (Hospital Universitario Puerta de Hierro-Madrid, España); Pózzoli, N.; Bender, C.; Deco, C.; Hernández, A. (Universidad Nacional de Rosario, Argentina). Ampliación de las capacidades de recuperar la información de la web a partir de una historia clínica electrónica. Publicado en Actas del VI Congreso Nacional de Informática de la Salud - INFORSALUD 2003, pp. 59-65. Madrid, 2-4 de abril de 2003.

[Plüss, 2004-2006] Proyecto de Investigación y Desarrollo (ING81) "Tecnologías Middleware e Internet: búsqueda asistida de evidencia clínica en medicina". Director del Proyecto: Dr. Jorge Plüss. Departamento de Sistemas e Informática, Facultad de Ciencias Exactas, Ingeniería y Agrimensura, Universidad Nacional de Rosario. 2004-2006.

[Pollitt, 1988] A. S. Pollitt. MenUSE for medicine: End user browsing and searching of MEDLINE via the MeSH thesaurus. In RIAO '88: User-oriented content based text and image handling, pages 547--573, Cambridge, MA, MIT, 1988.

[Ponce et al., 2008] Adrián Ponce, Claudia Deco, Cristina Bender. Proposal of an ontology based web search engine. En Proceedings del Workshop de Bases de Datos en el marco del XIV Congreso Argentino de Ciencias de la Computación, CACIC 2008. Chilecito, Argentina, octubre 2008.

[Porter, 1980] Porter, M. An Algorithm for Suffix Stripping. Program, 14:130–137, 1980.

[Porter, 2001] M. F. Porter. “Snowball, a Language for Stemming Algorithms”, October 2001: http://snowball.tartarus.org/texts/introduction.html

[Resnick et al., 1997] Resnick P. and Varian H. Recommender Systems. In Communications of the ACM, pp. 56-58, 1997.

[Romero et al., 2007] Romero, C., Ventura, S., Delgado, J. and de Bra, P. Personalized Links Recommendation Based On Data Mining in Adaptive Educational Hypermedia Systems. Second European Conference on Technology Enhanced Learning (EC-TEL 2007). Crete, Greece, 2007.

[Salton, 1970] Salton, G. Automatic Processing of Foreign Language Documents. Journal of American Society for Information Sciences, 21:187–194, 1970.

[Salton, 1983] Salton, G. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983.

[Sangoi Pizzato et al., 2003] Sangoi Pizzato, L., Strube de Lima, V. Evaluation of a Thesaurus-Based Query Expansion Technique. PROPOR’2003. Faro, Portugal, June 26-27, 2003.

[Savoy, 1999] Savoy, J. A Stemming Procedure and Stopword List for General French Corpora. Journal of the American Society for Information Science, 50:944–952, 1999.

[Schinke et al., 1996] Schinke, R., Robertson, A., Willet, P., Greengrass, M. A stemming algorithm for Latin text databases. Journal of Documentation, 52:172–187, 1996.

[SemanticWeb] http://www.semanticweb.org/. Consultado el 08-09-05.

[Shivakumar et al., 1998] Shivakumar, N., García Molina, H. Finding near-replicas of documents on the web. In Workshop on the Web Databases, Valencia, Spain, March 1998.

[Silberschatz et al., 1998] Silberschatz, A., Korth, H. Fundamentos de bases de datos. 3ra. ed. España. McGraw-Hill. 1998.

[Soderland, 1999] Soderland, S. Learning IE Rules for Semiestructured and Free Text. Machine Learning, 1999

[Solana, 2005-2008] Proyecto de Investigación y Desarrollo: "INFOSUR: Investigación y Desarrollo". Facultad de Humanidades y Artes de la Universidad Nacional de Rosario. Directora: Dra. Zulema Solana. 2005-2008.

[Studer, 1998] Studer S, Benjamins R., Fensel D. Knowledge Engineering: Principles and Methods. Data and Knowledge Engineering, vol. 25, pp. 161-197, 1998.

[Soonthornphisaj et al., 2006] Soonthornphisaj, N., Rojsattarat, E. and Yimngam, S. Smart E-Learning Using Recommender System. Lecture Notes in Computer Science. Springer Berlin, Heidelberg, Volume 4114. Computational Intelligence pp 518-523. 2006

[Terveen et al., 2001] Terveen, L. and Hill, W. Beyond Recommender Systems: Helping People Help Each Other. In Carroll, J. ed., HCI in the New Millennium. Addison Wesley, 2001.

[Tsikrika, 2001] Tsikrika, T, Information Retrieval, lecture, Queen Mary University of London, 2001.

[Voorhees, 1998] Voorhees, E., Using Wordnet for Text Retrieval, in Fellbaum C. "WordNet, an electronic Lexical Database", Mit Press. 1998.

[Vossen, 1998] Vossen, P. Introduction to EuroWordNet. Computers and the Humanities, Special Issue on EuroWordNet, 1998.

[W3C 1999] http://www.w3.org/TR/1999/REC-xslt-19991116 Recomendaciones del Consorcio W3C sobre XSLT. Consultado el 11-06-05.

[W3C] Extensible Markup Language (XML), Recomendaciones del Consorcio WWW. http://www.w3.org/TR/REC-xml. Consultado el 11-06-05.

[Welty et al., 2000] Welty, C., Jenkins, J. Untangle: a new ontology for card catalog systems. In Henry Kautz and Bruce Porter, eds., Proceedings of AAAI-2000: The National Conference on Artificial Intelligence. AAAI Press. July, 2000.

[Wiley, 2002] Wiley, D. Connecting Learning Objects to Instructional Design Theory: A definition, a metaphor, and a taxonomy. In D. A. Wiley (ed.) "Instructional Use of Learning Objects". Editorial Association for Instructional Technology. 2002.

[Zaiane , 2002] Zaiane, O.R. Building a recommender agent for e-learning systems. Proceedinds of International Conference on Computers in Education, pp: 55-59, 2002.

Apéndice 1: Relevamiento de Diccionarios Multilinguales y Tesauros disponibles en la Web

- ***Diccionarios multilinguales***

- **Foreignword.com** (http://www.foreignword.com/).
  Contiene diccionarios y herramientas de traducción. El traductor facilita la búsqueda mediante un sistema de búsqueda por palabras. Contiene varios idiomas, entre ellos el español.

- **Diccionarios.com** (http://www.diccionario.com/).
  Permite traducir términos del castellano al inglés, francés, italiano, catalán, euskera, y alemán. Contiene un diccionario general de lengua española y de sinónimos.

- **AllWords.com** (http://www.allwords.com/).
  Es un traductor virtual que transforma a alemán, español, francés, holandés inglés e italiano.

- **Rivendell's Machine Translation Dictionary**
  (http://rivendel.com/~ric/resources/translator.html)
  Traduce cualquier término al alemán, español, francés, italiano e inglés.

- **Online English to Spanish to English Dictionary** (www.freedict.com/onldict).
  Diccionario multilingüe de inglés con otros idiomas, incluido el español.

- **Babylon, Diccionario y Traductor** (http://www.babylon.com/)
  Diccionario multilingüe. Puede usarse en línea o puede ser descargado.

- **WordReference.com** (http://wordreference.com/).
  Traductor en línea multilingüe que traduce los idiomas español, alemán, italiano y francés al inglés y viceversa.

- **The Online dictionary** (http://dictionary.lezlisoft.com/dictionary/).
  Traductor virtual que traduce simultáneamente a alemán, español, húngaro e inglés.

- **Proyecto ARTFL** (http://humanities.uchicago.edu/forms_unrest/FR-ENG.html)
  Vocabulario constituido por 75.000 términos del francés y el inglés. Realizado por la Universidad de Chicago.

- ***Diccionarios multilinguales especializados***

- **Diccionario técnico textil** (www.textiledictionary.com/).

  Este diccionario contiene más de diez mil palabras y expresiones en cuatro idiomas (español, inglés, francés, alemán) referentes al sector textil, desde los diversos tipos de tejido hasta los procesos industriales aplicables a los mismos.

- **Eurodicautom** (http://europe.eu.int/eurodicautom/login.jsp)

  Es el diccionario multilingüe del Consejo Europeo y del Parlamento Europeo (alemán, danés, español, finlandés, holandés, inglés, italiano, portugués y sueco).

- **Fishbase Glossary** (http://www.fishbase.org/)

  Contiene definiciones en inglés, francés, español, y portugués de términos relacionados con disciplinas (ictiología, taxonomía, ecología, conservación, etc) vinculadas a las familias de peces.

- **ILOTERM** (http://www.ilo.org/iloterm/)

  Diccionario multilingual especializado de la International Labour Organization, contiene terminología social y laboral en los idiomas inglés, español, francés, alemán, ruso, chino y arábigo.

- ***Tesauros especializados***

Bellas Artes

- Art & Architecture Thesaurus Browser

http://www.getty.edu/research/tools/vocabulary/aat/index.html

- Thesaurus for Graphic Materials I: Subject Terms

http://www.loc.gov/pmei/lexico?usr=pubop=sessioncheck&db=TGM_I

- Thesaurus for Graphic Materials II: Genre and Physical Characteristic Terms

http://www.loc.gov/pmei/lexico?usr=pub&op=sessioncheck&db=TGM_II

Biblioteconomía y Documentación

- ASIS Thesaurus of Information Science

http://www.asis.org/Publications/Thesaurus/isframe.htm

- Dewey Decimal Classification - WWlib Browse Interface

http://www.scit.wlv.ac.uk/wwlib/browse.html

- Library of Congress Classification (LCC)

http://lcweb.loc.gov/catdir/cpso/lcco/lcco.html

Biomedicina

- CATIE Thesaurus

http://www.catie.ca/thesaurus.nsf/

- Medical Subject Headings (MeSH)

http://www.nlm.nih.gov/mesh/meshhome.html

- Tesauro de Ingeniería Sanitaria y Ambiental REPIDISCA

www.cepis.ops-oms.org/eswww/proyecto/repidisc/publica/tesauro/tesauro/tesaint

- Tesauro de Recursos Humanos en Salud

http://www.americas.health-sector-reform.org/sidorh/documentos/hsr2esp.html

- Tesauro sobre Reforma del Sector Salud

http://www.americas.health-sector-reform.org/spanish/clh2.htm

- The Alcohol and Other Drug (AOD) Thesaurus

http://etoh.niaaa.nih.gov/AODVol1/Aodthome.htm

- Thesaurus of Parasitology

http://www.personal.kent.edu/%7Eslis/zeng/template/thesauri/miller/tp.htm

Ciencias biológicas

- Aquatic Sciences & Fisheries Thesaurus

www.csa.com/htbin/ccfdisp.cgi?fn=/wais/data/thes/asfithes.ccf&sl=A&fmt=5&ldtag=TR

- Life Sciences Thesaurus

http://www.csa.com/edit/lscthes.html

- Tesauro de Agricultura Urbana

www.cepis.ops-oms.org/eswww/proyecto/repidisc/publica/tesauro/agri/tesauro.html

- Tesauros (CINDOC)

http://bdcsic.csic.es:8084/TESA/BASIS/tesa/carga/docu/SAI

Ciencias de la educación

- ERIC Thesaurus

http://searcheric.org/

- European Education Thesaurus – EET:

http://www.eurydice.org/TeeForm/frameset_en.HTM

- Humanities And Social Science Electronic Thesaurus

http://155.245.254.46/services/zhasset.html

- Tesauro de Educación: DAP

http://www.ucm.es/info/DAP/tesauro.htm

- Wordsmyth: The Educational Dictionary-Thesaurus

http://www.wordsmyth.net/

Derecho

- Eurovoc

http://europa.eu.int/celex/eurovoc/index.htm

- Global Legal Information Network (GLIN) Thesaurus

http://www.loc.gov/pmei/lexico?usr=pub&op=sessioncheck&db=GLIN

- Humanities And Social Science Electronic Thesaurus

http://155.245.254.46/services/zhasset.html

- Tesauro de la Materia Laboral

www.poder-judicial.go.cr/salasegunda/jurisprudencia/indice-tesauro-laboral-a.htm

- Tesauro de la Materia Familia y Civil

www.poder-judicial.go.cr/salasegunda/jurisprudencia/indice-tesauro-famciv-a.htm

Ecología y medio ambiente

- AGRIFOREST

http://wwwdb.helsinki.fi/triphome/agri/agrisanasto/Welcomeng.html

- GEneral Multilingual Environmental Thesaurus

http://www.mu.niedersachsen.de/cds/etc-cds_neu/library/select.html

- INFOTERRA Thesaurus Database

http://p5uni.ii.pw.edu.pl/envoc/

- Tesauro de Agricultura Urbana

www.cepis.ops-oms.org/eswww/proyecto/repidisc/publica/tesauro/agri/tesauro.html

- Tesauro de Medio Ambiente

http://medioambiente.comadrid.es/wwwhtm/residuos/cindoc/Tesauro/Med_amb.htm

- Umweltthesaurus / Environmental Thesaurus

http://udk.bmu.gv.at/

Economía, gestión y finanzas

- Humanities And Social Science Electronic Thesaurus

http://155.245.254.46/services/zhasset.html

- Le Thesaurus de Delphes

http://www.infomediatheque.ccip.fr/ccipdie/produits/thesaurus.htm

- OECD Macrothesaurus - HTML Version

http://info.uibk.ac.at/info/oecd-macroth/

- Tesauros (CINDOC)

http://bdcsic.csic.es:8084/TESA/BASIS/tesa/carga/docu/SAI

Física y astronomía

- PACS

http://www.aip.org/pacs/

- The Astronomy Thesaurus

http://msowww.anu.edu.au/library/thesaurus/

Geografía

- Feature Type Thesaurus

http://alexandria.ucsb.edu/%7Elhill/html/index.htm

- Tesauros (CINDOC)

http://bdcsic.csic.es:8084/TESA/BASIS/tesa/carga/docu/SAI

- Thesaurus of Geographic Names

http://www.getty.edu/research/tools/vocabulary/tgn/index.html

Informática

- Tesauro de Redes de Ordenadores

http://www.um.es/%7Egtiweb/fjmm/tesauro/

- Tesauro de Términos Informáticos

http://members.es.tripod.de/hv1102/tesauro.html

Ingeniería

- Canadian Thesaurus of Construction Science and Technology

http://www.nrc.ca/irc/thesaurus/ctcst-search-form.html

- NASA Thesaurus

http://www.sti.nasa.gov/thesfrm1.htm

- Tesauro de Ingeniería Hidráulica

http://hispagua.cedex.es/Grupo1/Tes_hidro/Tesauro.htm

- Tesauro de Ingeniería Sanitaria y Ambiental REPIDISCA

www.cepis.ops-oms.org/eswww/proyecto/repidisc/publica/tesauro/tesauro/tesaint

Lengua y literatura

- Humanities And Social Science Electronic Thesaurus

http://155.245.254.46/services/zhasset.html

- Lexical FreeNet: Connected Thesaurus

http://www.lexfn.com/

- Merriam Webster Thesaurus

http://www.m-w.com/thesaurus.htm

- SIGNUM

http://www.lenguaje.com/Tesauro/Default.htm

Multidisciplinarios

- Dewey Decimal Classification - WWlib Browse Interface

http://www.scit.wlv.ac.uk/wwlib/browse.html

- Humanities And Social Science Electronic Thesaurus

http://155.245.254.46/services/zhasset.html

- Library of Congress Classification (LCC)

http://lcweb.loc.gov/catdir/cpso/lcco/lcco.html

- UNESCO Thesaurus

http://www.ulcc.ac.uk/unesco/index.htm

Psicología y psiquiatría

- Fachgebärdenlexikon Psychologie

http://www.sign-lang.uni-hamburg.de/Projekte/PLEX/start.htm

- Humanities And Social Science Electronic Thesaurus

http://155.245.254.46/services/zhasset.html

Sociología

- Eurovoc

http://europa.eu.int/celex/eurovoc/index.htm

- Humanities And Social Science Electronic Thesaurus

http://155.245.254.46/services/zhasset.html

- Population Multilingual Thesaurus

http://www.cicred.ined.fr/thesaurus/integral/

- Sociology Thesaurus

http://www.csa.com/htbin/ccfdisp.cgi?fn=/wais/data/thes/asfithes.ccf&sl=A&fmt=5&

- Thesaurus of Sociological Indexing Terms

http://www.csa.com/edit/sociothes.html

Apéndice 2: Problemas de la traducción de la consulta en la búsqueda de información multilingüe

La Recuperación de Información Multilingüe trata el problema de encontrar documentos que están escritos en otros idiomas, distintos al idioma de la consulta. Si se desea recuperar documentos en otro idioma, es necesario efectuar una traducción de la consulta para realizar la búsqueda en dicho idioma. Este proceso no es simple debido a la complejidad semántica del vocabulario. La necesidad de realizar búsquedas multilingües es un hecho, y la demanda de este tipo de búsquedas aumenta con el crecimiento de la Web.

En este apéndice se presenta el problema de la búsqueda de información multilingüe, con especial atención a los distintos recursos lingüísticos que se pueden utilizar, y los problemas que se presentan en la traducción de la consulta. Además, se presentan los resultados de la experimentación realizada para evaluar algunos diccionarios multilingües disponibles en línea, para traducciones entre los idiomas español, inglés y francés.

A.2.1. Traducción de la consulta

El problema en una búsqueda multilingüe de información es que los idiomas de la consulta y de los documentos son distintos. Por lo tanto, es necesario efectuar una traducción para poder realizar una búsqueda en la que tanto la consulta como los documentos se encuentren en el mismo idioma.

[Salton, 1970] planteó por primera vez el problema de encontrar documentos escritos en un idioma diferente al de la consulta. Propuso la utilización de un tesauro bilingüe alemán-inglés. Los resultados obtenidos fueron similares a los de una búsqueda monolingüe, debido a que el tesauro utilizado había sido construido manualmente. De esta forma la correspondencia entre los términos entre ambos idiomas era perfecta y no existía ambigüedad en los términos de búsqueda.

En el problema de la recuperación de información multilingüe, la traducción

de la consulta es la opción más frecuente, porque su costo computacional es menor al costo de traducir los documentos.

Los tres problemas principales para automatizar la traducción de la consulta, según [Grefenstette, 1998], son: saber cómo un término escrito en un idioma puede ser expresado en otro idioma; decidir cuáles de las posibles traducciones de cada término son las adecuadas en un contexto dado; y saber cómo medir la importancia de las diferentes traducciones que se consideran adecuadas. Estos problemas son compartidos por los sistemas de traducción automática y los sistemas de recuperación de información multilingüe.

El refinamiento propuesto en esta tesis puede ser extendido a la recuperación de información multilingüe si en la etapa de *expansión*, se utilizan *recursos multilingües* para traducir los términos originales a otros idiomas, realizando así una expansión multilingüe de la consulta.

Para realizar la traducción automática de la consulta se pueden utilizar recursos tales como *diccionarios multilingües* y *tesauros multilingües*.

Un *diccionario multilingüe* puede ser general o especializado. Un ejemplo de diccionario multilingüe general es EuroWordNet [Vossen, 1998], que es una base de datos multilingüe con redes de palabras para varios de los idiomas europeos: holandés, italiano, español, alemán, francés, checo y estonio. Está basado en el diccionario WordNet. Los idiomas están interconectados de forma que se puede ir de palabras en un idioma a sus palabras equivalentes en cualquiera de los otros idiomas.

Un *tesauro multilingüe* sobre un área del conocimiento permite la traducción de términos específicos de ese dominio que quizá no puedan encontrarse en un diccionario. Los tesauros multilingües son recursos diseñados específicamente para la recuperación multilingüe de información. Un ejemplo de este tipo de tesauro sobre el dominio médico es UMLS (Unified Medical Language System), que es el Sistema Unificado de Terminología Médica de la Biblioteca Nacional de Medicina de Estados Unidos [National Library of Medicine, 1997]. Otro ejemplo, de tesauro multilingüe general es EuroVoc, de la Comunidad Europea, que abarca nueve idiomas [EuroVoc, 1995].

Otra posibilidad para la traducción de la consulta es el uso de *programas de*

traducción automática. En consultas formadas por frases, el uso de estos programas produce una mejora en la desambiguación, frente al uso de diccionarios que traducen palabras aisladas. Esto se debe a que los sistemas de traducción automática consideran la estructura sintáctica del texto.

Sin embargo, el uso de un diccionario como recurso en la traducción automática de la consulta presenta problemas, tales como los siguientes:

- Los términos específicos o técnicos, propios de un área del conocimiento, pueden no existir en un diccionario de uso general. Para dominios específicos del conocimiento se logran mejores resultados si se utilizan diccionarios especializados.
- En un diccionario pueden no estar todas las variantes morfológicas de una palabra. Este problema se soluciona utilizando la técnica de stemming, llevando la palabra no encontrada a su forma raíz y buscando ésta en el diccionario.
- Muchos diccionarios no tienen traducciones para los sustantivos propios.
- Una palabra en un idioma, puede tener varias traducciones distintas en otro idioma. Por ejemplo, el término “investigación” en español, se traduce al inglés como “research” o “investigation”. El primero se utiliza para la investigación científica y el segundo para la policial. Entonces, para decidir cuál es la traducción adecuada, debe contemplarse el contexto. Este es un problema complejo, ya que se debe automatizar la desambiguación de la traducción.
- Muchos diccionarios no tienen traducciones para conceptos formados por varias palabras, es decir por frases. La traducción de cada término por separado puede llevar a un error en la traducción del concepto.
- Los distintos diccionarios no tienen criterios uniformes para la traducción de los conceptos, ya que distintos diccionarios traducen una misma palabra de distinta forma.

Es de interés en esta tesis analizar algunos diccionarios multilingües disponibles en línea a fin de evaluar qué problemas presentan para contemplarlos en la automatización de la expansión de la consulta. En la siguiente sección, se presenta la experimentación realizada.

A.2.2. Experimentación

El objetivo de las experiencias fue evaluar algunos diccionarios multilingües, disponibles en línea, para las traducciones entre los idiomas español, inglés y francés. Para esto, se utilizaron los siguientes recursos:

- Systran (tr.voila.fr).
- Reverso (www.elmundo.es/traductor/): traductor del diario El Mundo de España.
- El servicio de SDL internacional (www.freetranslation.com/). Este servicio no ofrece la traducción del español al francés.
- Wordlingo (www.worldlingo.com/en/products_services/worldlingo_translator).

Los resultados de estas experiencias se encuentran en las Tablas A.2.1 y A.2.2. En la Tabla A.2.1 se muestran las traducciones del español al inglés. En la Tabla A.2.2 se presentan las traducciones del español al francés. En ambos casos se utilizó el mismo grupo de términos.

Término en Español	Traducción de Systran	Traducción de Reverso	Traducción de SDL	Traducción de Wordlingo
Alemania	Germany	Germany	Germany	Germany
Almohada	Pillow	Pillow	Pillow	Pillow
Anglosajón	Anglo-saxon	Anglo-saxon	Anglo-saxon	Anglo-saxon
Arreglo	Adjustement	Arrangement	I arrange	Adjustment
Ayuda	Aid	Help	It helps	Aid
Bandeja	Tray	Tray	Tray	Tray
Base De Datos	Data base	Base of information	Database	Data base
Basto	Coarse	Pack-saddle	I suffice	Coarse
Bujía	Spark plug	Candlestick / Spark plug	Sparkplug	Spark plug
Callo	Callus	Corn	I silence	Callus
Camboya	Cambodia	Cambodia	Cambodia	Cambodia
Cisne	Swan	Swan	Swan	Swan
Comida	Food	Food	Food	Food
Erizo	Sprocket wheel	Hedgehog	I bristle	Sprocket wheel
Falta	Lack	Lack / Mistake	It lacks	Lack
Ginebra	Geneva	Geneva	Geneva	Geneva
Guardarropa	Wardrobe	Wardrobe	Coat room	Wardrobe
Hamaca	It swings	Hammock	Hammock	Hammock
Lamento	Moan	Lament	Lament	Moan
Loco	Crazy person	Madman	Crazy	Crazy person
Loza	Stoneware	Crockery	China	Stoneware
Matriz	Matrix	Counterfoil	Headquarters	Matrix
Mesa	It pulls	Table	Table	Table
Móvil	Movable	Mobile	Mobile	Moving body
Pánico	Panic	Panic	Panic	Panic
Pekin	The beijing	Pekin	Pekin	The beijing
Remera	Rower	Remere	Oarswoman	Rower
Ruido	Noise	Noise	Noise	Noise
Telaraña	Spiderweb	Spiderweb	Web	Spiderweb
Tocino	Bacon	Bacon	Bacon	Bacon
Ultra Rápido	Extreme express	Ultra rapid	Right-wing fast	Extreme express
Uso	Use	Use	Use	Use
Zorra	Vixen	Fox	Foxy	Vixen

Tabla A.2.1: Traducciones entre el español y el inglés

Término en Español	Traducción de Systran	Traducción de Reverso	Traducción de Wordlingo
Alemania	L'Allemagne	L'Allemagne	L'Allemagne
Almohada	Oreiller	Oreiller	Oreiller
Anglosajón	Anglo-saxon	Anglo-saxon	Anglo-saxon
Arreglo	Ajustement	Entente	Ajustement
Ayuda	Aide	Aide	Aide
Bandeja	Plateau	Plateau	Plateau
Base de datos	Base de données	Base de données	Base de données
Basto	Brut / Je suffis	Bât	Je suffis
Bujía	Bougie	Chandelier / Bougie	Bougie
Callo	Calus	Grain / Maïs	Calus
Camboya	Le Cambodge	Le Cambodge	Le Cambodge
Cisne	Cygne	Cygne	Cygne
Comida	Repas	Alimentation	Repas
Erizo	Hérisson	Hérisson	Hérisson
Falta	Manque	Manque / Erreur	Manque
Ginebra	Genève	Genève	Genève
Guardarropa	Guardarropa	Garde-robe	Guardarropa
Hamaca	Hamac	Hamac	Hamac
Lamento	Je regrette	Lamenter	Je regrette
Loco	Fou	Fou	Fou
Loza	Faïence	Poterie	Faïence
Matriz	Matrice	Souche	Matrice
Mesa	Table	Table	Table
Móvil	Mobile	Portable	Raison
Pánico	Panique	Panique	Panique
Pekin	Pekin	Pékin	Pekin
Remera	Rémige	Resimple	Rémige
Ruido	Bruit	Bruit	Bruit
Telaraña	Toile d'araignée	Spiderweb	Toile d'araignée
Tocino	Lard	Bacon	Lard
Ultra rápido	Ultra rapide	Ultra rapide	Ultra rapide
Uso	Utilisation	Utilisation	Utilisation
Zorra	Renard	Renard	Renard

Tabla A.2.2: Traducciones entre el español y el francés

En estas tablas, se observa que el término *Basto*, que puede corresponder a un sustantivo o a un verbo conjugado, es traducido por Systran como sustantivo y como verbo, para las traducciones al francés. Pero Reverso lo traduce como sustantivo solamente y Wordlingo lo traduce como verbo solamente. En las traducciones al inglés, sólo SDL lo traduce como verbo, el resto lo traduce como sustantivo.

Los términos *Hamaca*, *Ayuda*, *Arreglo*, *Falta*, *Uso* y *Callo*, que pueden corresponder tanto a un verbo conjugado como a un sustantivo, son traducidos por todos los traductores al francés como sustantivo. Sin embargo, en el caso de *Lamento*, la mayoría de los traductores analizados lo traducen como verbo. En sus traducciones al inglés, Systran es el único que considera a *Hamaca* como verbo; y SDL es el único que considera a *Ayuda*, *Arreglo*, *Falta*, y *Callo* como verbos conjugados.

Respecto a los sustantivos propios, Reverso en su traducción al inglés los interpreta como tales si están escritos en mayúsculas. Así, *Ginebra* lo traduce como *Geneva*, pero *ginebra* lo traduce como *gin*. Si un sustantivo propio se ingresa en minúsculas, y no corresponde a un sustantivo común, ni Reverso ni Systran lo traducen.

En sus traducciones del español al francés del término *Pekín*, tanto Systran como Wordlingo, omiten la acentuación de la letra “e”, lo que es un error en francés.

Reverso traduce *Telaraña* al francés como *Spiderweb*. Esto es llamativo, porque ninguna de las dos componentes de esta palabra (*spider* y *web*) son de origen francés. Sin embargo, con Reverso, *Spiderweb* no es traducida al español ni al inglés.

Se ha observado además que en algunos casos, que se detallan a continuación, se presenta el problema de que la traducción no es bidireccional. En las traducciones entre el español y el inglés realizadas por Reverso, se advirtió que:

- *Matriz* lo traduce al inglés como *Counterfoil*. *Counterfoil* lo traduce al español como *Talón*. *Talón* lo traduce al inglés como *Heel*. Sin embargo, el término *Matrix* lo traduce al español como *Matriz*.

- *Callo* lo traduce al inglés como *Corn*. *Corn* lo traduce al español como *Grano*. *Grano* lo traduce al inglés como *Grain*. *Callus* lo traduce al español como *Callo*
- *Basto* es traducido al inglés como *Pack-saddle*. *Pack-saddle* es traducido al español como *Albarda*.

Y en las traducciones entre el español y el francés realizadas por Reverso, se observó que:

- *Arreglo* lo traduce al francés como *Entente*. *Entente* lo traduce al español como *Armonía*. *Armonía* lo traduce al francés como *Harmonie*.
- *Loza* lo traduce al francés como *Poterie*. *Poterie* lo traduce al español como *Alfarería*. Sin embargo, *Faïence* lo traduce al español como *Loza*.
- *Callo* lo traduce al francés como *Grain* (Maïs). *Grain* lo traduce al español como *Grano*. *Cal*, traducida al español, da como resultado *Callo*. *Durillon*, traducido al español, da como resultado *Callosidad*.
- *Basto* es traducido al francés como *Bât*. Pero *Bât* no es reconocido para traducirlo al español.
- *Móvil* es traducido al francés como *Portable*. *Portable* es traducido de idéntica forma al español. Sin embargo, *Mobile* es también traducido al español como *Móvil*.

En las traducciones entre el español y el inglés realizadas por Wordlingo, se observó que:

- *Bujía* es traducida al inglés como *Spark plug*. *Spark plug* es traducido al español como *Chispa Enchufe*. *Chispa* es traducido al inglés como *Spark*. *Enchufe* es traducido como *Fit*. *Fit* es traducido al inglés como *Ajuste*. Sin embargo, *Sparkplug* (todo junto) sí es traducido al español como *Bujía*.
- *Comida* es traducido al inglés como *Food*. *Food* es traducido al español como *Alimento*. *Meal* es traducido al español como *Comida*.
- *Lamento* es traducido al inglés como *Moan*. *Moan* es traducido al español como

Quejido. *Quejido* es traducido al inglés como *Complaint*. *Complaint* es traducido al español como *Queja*.

- *Almohada* es traducido al inglés como *Pillow*. *Pillow* es traducido al español como *Almohadilla*. *Almohadilla* es traducido al inglés como *Pad*. *Pad* es traducido al español como *Cojín*. *Cojín* es traducido al inglés como *Cushion*. *Cushion* es traducido al español como *Amortiguador*.

Con sustantivos compuestos, también se presenta el problema de la traducción bidireccional. En este sentido, se observó que:

- SDL traduce *Guardarropas* como *Coat room*. *Coat room* es traducida al español como *Revista el espacio*. Sin embargo, *Wardrobe* es traducida al español como *Guardarropa.*
- SDL traduce *Telaraña* como *Web*. *Spiderweb* también es traducido al español como *Telaraña*. Sin embargo, Systran y Wordlingo traducen *Web* como *Tela*. Reverso no traduce *Web* al español.
- Reverso traduce *Base de datos* como *Base of information*. *Base of information* es traducida al español como *Base de información*. Sin embargo, Reverso traduce al español la palabra inglesa *Database* como *Base de datos*.
- Systran traduce *Ultra rápido* al inglés como *Extreme express*. Pero, *Extreme express* lo traduce al español como *Extremo expreso*. Sin embargo, traduce la palabra inglesa *Ultrarapid* al español como *Ultrarrápido*. Esta última palabra no es de existencia reconocida por la Real Academia Española.

Por todos estos problemas, la utilización de un diccionario como único recurso de traducción reduce la efectividad de las búsquedas multilingües.

Diversos trabajos, como los de [Hull & Grefenstette, 1996] y [Ballesteros & Croft, 1996], comprueban que si se sustituye cada término de la consulta por todas las traducciones ofrecidas por el diccionario, la efectividad se reduce entre un 40 y un 60%, respecto de la misma búsqueda realizada en un contexto monolingüe.

Con respecto a la polisemia, [Davis, 1997] propone utilizar la categoría

gramatical de las palabras de la consulta para elegir entre las posibles traducciones de los términos. Utilizando un diccionario bilingüe con información sobre la categoría gramatical para traducir las consultas, Davis comprobó que esta estrategia incrementaba en un 37% la precisión con respecto a la estrategia de sustituir cada término por todas las traducciones ofrecidas por el diccionario.

[Ballesteros & Croft, 1997] intentan mejorar la efectividad de las traducciones utilizando traductores de expresiones multipalabra. Con este tipo de recurso, las búsquedas fueron aproximadamente 150% más eficientes que aquellas en las que se tradujo cada palabra por separado.

[Pirkola, 1998] concluye que la traducción de la consulta escrita mediante una oración completa en lenguaje natural provee una mayor precisión que si la consulta está expresada con palabras aisladas y se traduce cada palabra por separado. Además, para la traducción experimentó varias formas de combinar dos diccionarios bilingües: uno de propósito general y otro específico del dominio. Comprueba así que los mejores resultados se obtenían al utilizar todas las distintas traducciones proporcionadas por ambos diccionarios.

En [Boughanem et al., 2002] se realiza una selección de las traducciones empleando las traducciones inversas, seleccionando sólo aquellas que pueden volver a traducirse al término de partida. Los resultados obtenidos en este trabajo muestran que esta estrategia puede ser más efectiva que otras más complejas, como la desambiguación de traducciones.

La interacción con el usuario es fundamental para solucionar estos problemas. Un sistema de búsqueda de información debe proporcionar al usuario la capacidad de expresar su necesidad de información en su propio idioma y ayudarlo a traducirla al idioma en el cual se encuentran los documentos. Para esto, el sistema puede utilizar un diccionario para traducir cada término de la consulta, permitiéndole al usuario, en el caso de términos ambiguos, seleccionar la traducción adecuada. A partir de esta selección, el sistema de búsqueda de información puede realizar una búsqueda automática.

En el caso de la traducción de frases pueden ocurrir que traducciones correctas no arrojen resultados. Por ejemplo, en el caso de *Enfermedad de Munchausen*, la traducción al inglés realizada por Systran es *Disease of Munchausen*.

Cuando se busca esta frase en Google, la búsqueda arrojó cero resultados. En cambio, si se utilizan *Muchausen disease* o *Munchausen's disease*, traducciones provistas por un usuario especialista en temas médicos, se obtuvieron 286 resultados y 150 resultados respectivamente. Con esto se ve la importancia de utilizar recursos especializados en cada área del conocimiento y no diccionarios o recursos generales, en el caso de búsquedas especializadas.

La frase *Polimialgia reumática* es traducida por Systran como *Rheumatic polimialgia*, que buscada en Google arroja 3 resultados. Sin embargo, la traducción correcta es *Rheumatic polymyalgia*, que buscada en Google arroja 472 resultados. Por lo tanto, una mala traducción (utilizar la traducción incorrecta *polimialgia* en lugar de la correcta *polymyalgia*), cuyo error pase inadvertido, puede llevar a obtener malas conclusiones, puesto que aún frases incorrectas arrojan algún tipo de resultados, lo que puede inducir a pensar que la traducción fue acertada y que en realidad no hay información abundante sobre eso en la Web.

Un enfoque distinto al presentado hasta aquí de traducción de la consulta, es la traducción de los documentos al idioma utilizado en la escritura de la consulta. Según [Dumais et al., 1996] y [Oard, 1998], este enfoque brinda traducciones más precisas porque se cuenta con información del contexto en el que se utilizan las palabras. Pero el problema que se presenta en este caso, es que el tiempo que lleva traducir los documentos es mucho mayor que el necesario para traducir la consulta.

A.2.3. Conclusiones sobre traducción de la consulta

En el globalizado mundo actual, la tecnología pone a disposición de quienes pueden acceder a ella una gran masa de documentos de infinitud de temáticas y entre los cuales se encuentran textos de altísimo valor. Estos textos pueden estar en un idioma distinto al utilizado para la consulta. La necesidad de realizar búsquedas multilingües es un hecho, y la demanda de este tipo de búsquedas aumenta con el crecimiento de la Web. La Recuperación de Información Multilingüe trata el problema de encontrar documentos que están escritos en otros idiomas, distintos al idioma de la consulta. Este proceso no es simple debido a la complejidad semántica

del vocabulario.

En este apéndice, se presentó el problema de la búsqueda de información multilingüe, con especial atención a distintos recursos lingüísticos que pueden utilizarse, y los problemas que se presentan en la traducción de la consulta.

En una búsqueda multilingüe de información, los idiomas de la consulta y de los documentos son distintos. Por lo tanto, es necesario efectuar una traducción para poder realizar una búsqueda en la que tanto la consulta como los documentos se encuentren en el mismo idioma. La traducción de la consulta es la opción más frecuente, porque su costo computacional es menor al costo de traducir los documentos. La traducción será de gran ayuda, a condición de que se trate de un trabajo de gran precisión. Sin embargo, los progresos logrados en la traducción automática de textos no logran poner a la misma en un pie de igualdad con la traducción humana, que sigue siendo, con mucho, más exacta y comprensible.

Para realizar la traducción automática se pueden utilizar recursos tales como diccionarios multilingües y tesauros multilingües. Otra posibilidad es el uso de programas de traducción automática. En consultas formadas por frases, el uso de estos programas produce una mejora en la desambiguación, frente al uso de diccionarios que traducen palabras aisladas. Esto se debe a que los sistemas de traducción automática consideran la estructura sintáctica del texto. Una tercera posibilidad es trabajar directamente con la consulta expresada en lenguaje natural. La traducción en este caso, provee una mayor precisión que si la consulta está expresada con palabras aisladas y se traduce cada palabra por separado.

Las experiencias realizadas tuvieron como objetivo evaluar algunos diccionarios multilingües, disponibles en línea, para las traducciones entre los idiomas español, inglés y francés. Los diccionarios utilizados fueron: Systran, Reverso, SDL y Wordlingo. De estas experiencias se ha observado que en algunos casos la traducción no es bidireccional. Otro problema que se presenta es que muchos diccionarios no tienen traducciones para conceptos formados por varias palabras, ni para los sustantivos propios, ni para términos específicos o técnicos. Además, una palabra puede tener varias traducciones distintas. En este caso, para decidir cuál es la traducción adecuada, debe contemplarse el contexto. Por todos estos problemas, la utilización de un diccionario como único recurso de traducción

reduce la efectividad de las búsquedas multilingües.

Una primera discusión sobre recursos multilingües se presentó en [Deco et al., 2006b]. Los resultados de las experiencias realizadas fueron publicados en [Deco et al., 2007].

La interacción con el usuario es fundamental para solucionar los problemas de traducción. El sistema puede utilizar un diccionario para traducir cada término de la consulta, permitiéndole al usuario, en el caso de términos ambiguos, seleccionar la traducción adecuada, y a partir de esta selección el sistema puede realizar una búsqueda automática.

En el refinamiento semántico propuesto en esta tesis, la desambiguación del término que realiza el usuario en la primera etapa de la preparación de la estrategia de consulta fija la acepción de interés del término. Esto permite mejorar el proceso de traducción de la consulta eligiendo la traducción correspondiente a la acepción de interés.

Apéndice 3: Prototipo

Descripción del prototipo

Este prototipo resuelve la generación de la estrategia de búsqueda a partir de conceptos ingresados por el usuario, siguiendo una secuencia de pasos.

La Figura A.3.1 muestra la pantalla inicial del prototipo. En la versión actual se requiere que el concepto a buscar se ingrese en inglés. Esta limitación se debe al uso de WordNet como recurso libre de la web.

Ingreso del concepto:

En la pantalla de la Figura A.3.1, se encuentra una caja de texto y por encima de ella una etiqueta indicando que es allí donde se debe ingresar –en inglés- la palabra que se desea buscar. A su derecha, el botón de búsqueda "search" tiene la función de enviar al sistema la palabra ingresada.

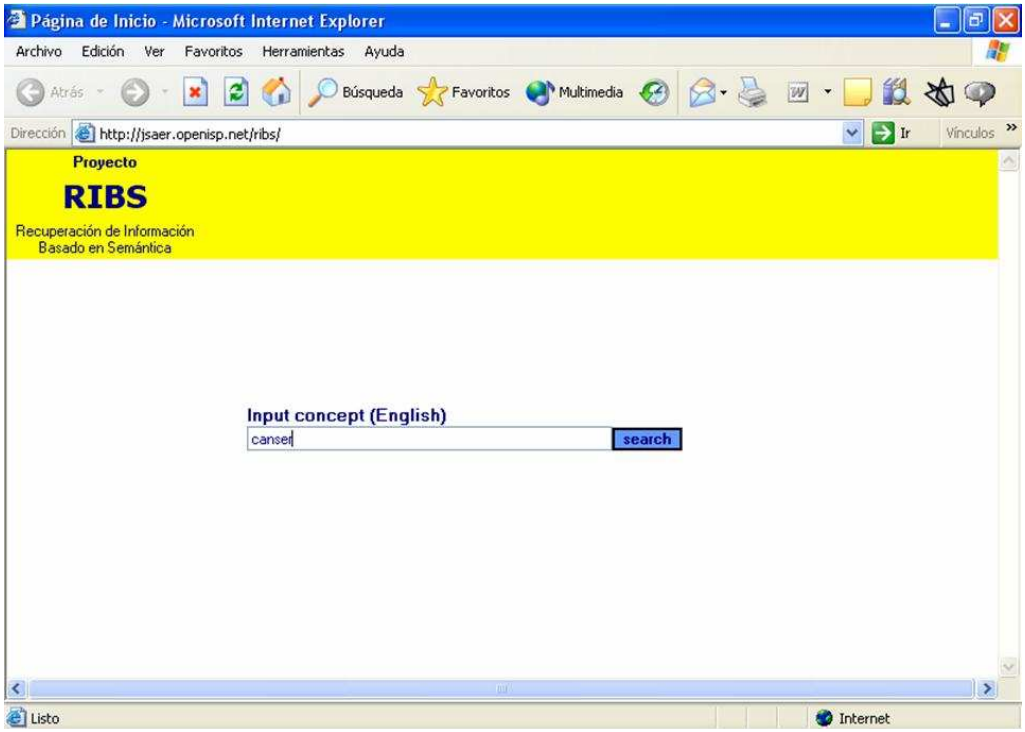


Figura A.3.1. Pantalla inicial del prototipo de refinador semántico

Corrección ortográfica:

Al oprimir el botón de búsqueda se presenta una segunda pantalla correspondiente al resultado de la corrección ortográfica. Para esta corrección se utiliza el recurso Google Web Service – Spelling Suggestion.

Si el término ingresado en la pantalla inicial, es considerado por el corrector ortográfico como inexistente o mal escrito, se le ofrece al usuario dos opciones para continuar el refinamiento. Una opción es continuar el proceso sobre el término ingresado originalmente por el usuario. La otra opción es continuar con un término cuya grafía es aproximada a la ingresada y que sí aparece como correcto según el corrector. Cada opción aparece como hipervínculo, como se muestra en la Figura A.3.2.

Esta posibilidad de continuar con el término original, y no con el corregido, se provee porque existen ciertos términos, como ser nombres propios, siglas, etc., que el corrector ortográfico no reconoce y sí pueden ser de interés para el usuario. Por ejemplo, si se ingresa el término “bender”, correspondiente a un apellido, el corrector sugiere como término bien escrito el término “vender”. El usuario en este caso puede decidir continuar con el término que él ingresó, si desea información relacionada con este apellido.

Si el término ingresado en primera instancia es correcto ortográficamente, aparece en esta pantalla un cartel invitando a continuar el proceso de refinamiento con dicha palabra.

Supongamos que el usuario intenta ingresar el término “cancer”, pero lo tipea en forma incorrecta, como “canser”. En la Figura A.3.2 se muestra la pantalla resultante del corrector ortográfico para este caso.

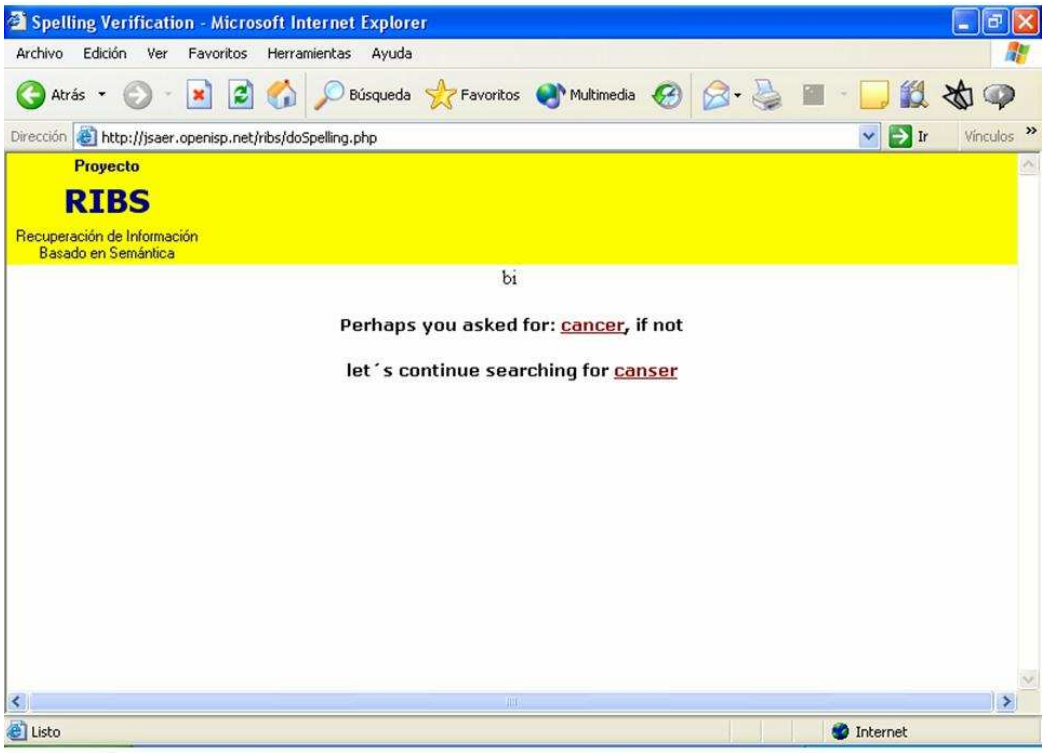


Figura A.3.2. Pantalla resultante de la corrección ortográfica para un término ingresado con errores.

En esta pantalla el usuario debe seleccionar cuál de los dos términos es de su interés: el término sugerido por el corrector o el término ingresado por el usuario. Consideremos que decide seleccionar el término corregido “cancer”.

Desambiguación:

Un término puede tener más de una acepción. En este paso del refinamiento, se le pide al usuario que desambigüe el término, eligiendo la acepción de su interés. Para esto se muestra al usuario una pantalla con las diferentes acepciones del término seleccionado en el paso anterior, si es que éste tiene más de una acepción. Para el prototipo se utiliza el recurso WordNet para mostrar las distintas acepciones.

Esta desambiguación que realiza el usuario permite continuar el proceso, en las etapas siguientes, con sólo aquellos términos que están vinculados conceptualmente con la acepción de interés del usuario.

Para continuar el usuario debe seleccionar la acepción de su interés. Las distintas acepciones aparecen como hipervínculos como se muestra en la Figura A.3.3.

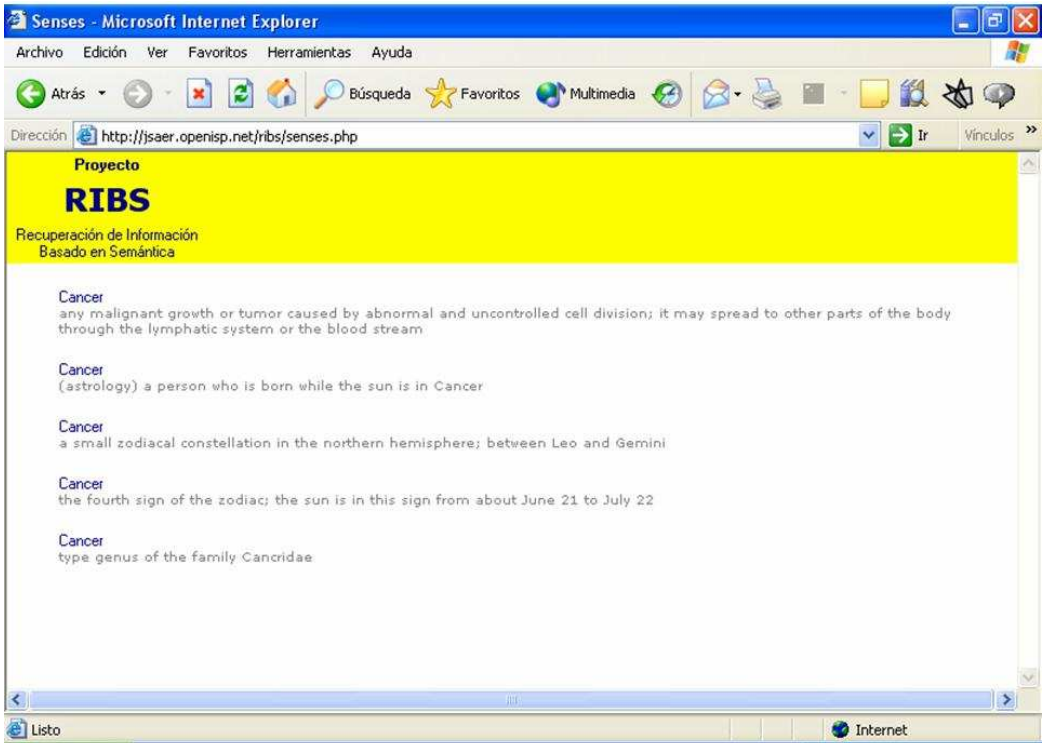


Figura A.3.3. Pantalla que muestra las distintas acepciones de un término.

Continuando con el ejemplo, en la Figura A.3.3, se muestra la pantalla que contiene las distintas acepciones del término “cancer”. En este ejemplo el usuario decide seleccionar la primera acepción correspondiente a la enfermedad.

Selección jerárquica:

Luego que el usuario ha optado por alguna de las acepciones del término, el prototipo despliega una nueva pantalla como la de la Figura A.3.4. En esta pantalla se muestra:

- el término: *cancer*,
- su hiperónimo: *malignant tumor*

- y por debajo una lista de sus hipónimos: *lymphoma*, *carcinoma* ...

todos ellos en forma de hipervínculo.

El usuario puede navegar por la jerarquía conceptual a través de estos hipervínculos. Si el usuario elige un hipónimo, se le muestra en pantalla la nueva jerarquía conceptual, donde el término original pasa a ser su hiperónimo y se muestran los hipónimos del nuevo término.

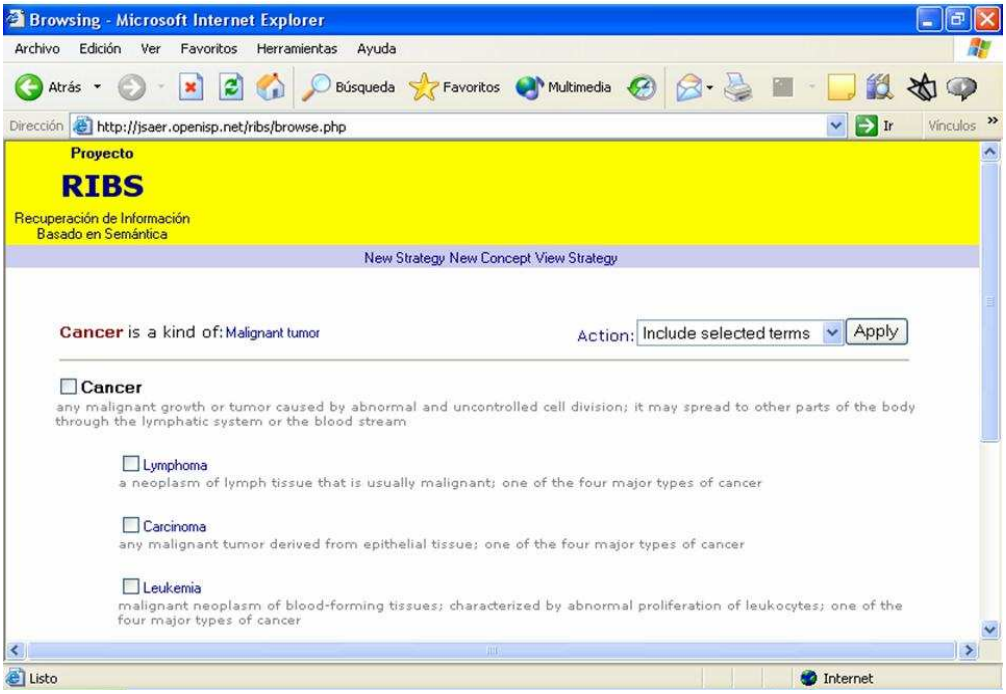


Figura A.3.4. Pantalla que muestra parte de la jerarquía conceptual de "cancer" en medicina.

Continuando con el ejemplo, en la Figura A.3.4, se muestra la pantalla que contiene la jerarquía conceptual del término "cancer" en su acepción médica. En esta figura se observa el término "malignant tumor", que es un hiperónimo de "cancer", y la lista de sus hipónimos. Si el usuario hace click en el hipervínculo del hipónimo "leukemia", se muestra la nueva jerarquía conceptual como se ve en la Figura A.3.5.

En lugar de elegir un hipónimo, el usuario podría elegir el hiperónimo. Es decir, puede ascender o descender en la jerarquía conceptual.

En la navegación por la jerarquía se mantiene la acepción del concepto elegida por el usuario en el paso de desambiguación. Es decir, no se vuelve a requerir al usuario que vuelva a desambiguar algún concepto relacionado jerárquicamente, si éste tiene más de una acepción.

A la izquierda del término y de cada hipónimo, se incluye un checkbox, que le permite al usuario seleccionar uno o más términos. Tildados los términos de interés, la opción *Include selected terms* del listbox situado en la parte superior derecha de la pantalla, permite incluirlos en la estrategia de búsqueda.

Los términos seleccionados se incorporan con el operador OR en la estrategia de búsqueda. Esto es transparente para el usuario.

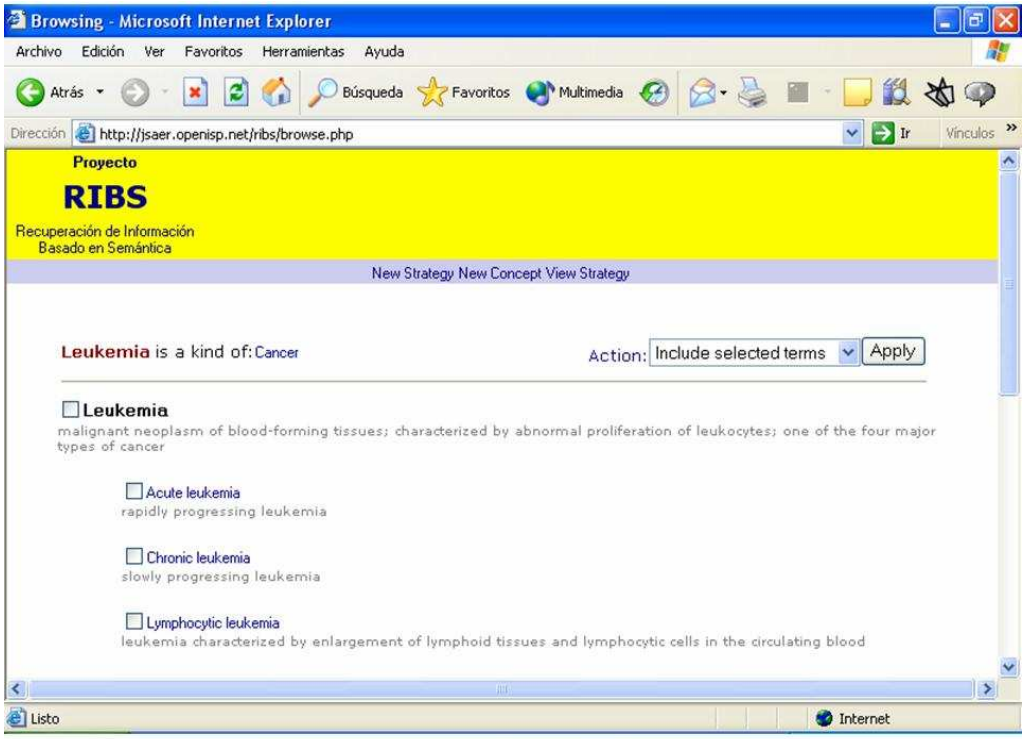


Figura A.3.5. Pantalla que muestra la jerarquía conceptual de "leukemia".

El usuario puede decidir continuar con el término de partida "cancer", o puede elegir uno, o más, términos específicos, tildando los hipónimos correspondientes. Un ejemplo de esto es que al usuario le interesen ciertos tipos de

cáncer. Podría así elegir “lung cancer”, “liver cancer” y “leukemia”; y seguir la búsqueda con estos términos.

El mecanismo para excluir términos de la estrategia de búsqueda es similar. Luego de tildar el término raíz o los términos específicos que desea descartar, el usuario elige la opción *Exclude selected terms* del listbox situado en la parte superior derecha de la pantalla.

Los términos seleccionados se incorporan con el operador NOT en la estrategia de búsqueda. Esto es transparente para el usuario.

Expansión semántica:

El o los términos de interés incluidos en el paso anterior se expanden automáticamente incorporando también sus sinónimos a la estrategia de búsqueda. Los sinónimos se agregan a la estrategia con el operador lógico OR. Por ejemplo, si el término seleccionado es “leukemia”, el refinador incorpora también los términos “leukaemia”, “leucaemia”, “cancer of the blood”.

Esto mismo ocurre al excluir un término. Se descartan también sus sinónimos, anteponiendo el operador NOT al OR lógico de los términos a excluir. Por ejemplo, si se desea excluir el término “lung cancer”, se excluye también su sinónimo: “carcinoma of the lungs”. Esto es transparente para el usuario.

Generación de la estrategia:

Todo el proceso descrito hasta este punto fue para un único concepto de partida. En general, una estrategia de búsqueda consiste de varios conceptos.

Como se observa en el Figura A.3.5, en la parte superior de la pantalla, existen las siguientes opciones: New Strategy, New Concept y View Strategy.

Una estrategia de búsqueda puede involucrar varios conceptos.

La opción *New Strategy*, cancela la estrategia actual y permite comenzar una nueva estrategia de búsqueda.

La opción *New Concept*, permite agregar un concepto a la estrategia actual. Es decir, le permite al usuario repetir el procedimiento descrito para un nuevo concepto correspondiente a su consulta actual. Si se elige esta opción, el prototipo vuelve a mostrar la pantalla que se muestra en la Figura A.3.1.

La opción *View Strategy*, muestra en pantalla la estrategia generada hasta el momento. Por ejemplo, para la consulta "quimioterapia utilizada en leucemia", la estrategia resultante es la mostrada en la Figura A.3.6. En esta versión del prototipo, la estrategia mostrada se puede editar, para permitirle al usuario modificar manualmente la estrategia si lo desea. Si se presiona el botón *Buscar en Yahoo* se envía la consulta al buscador.

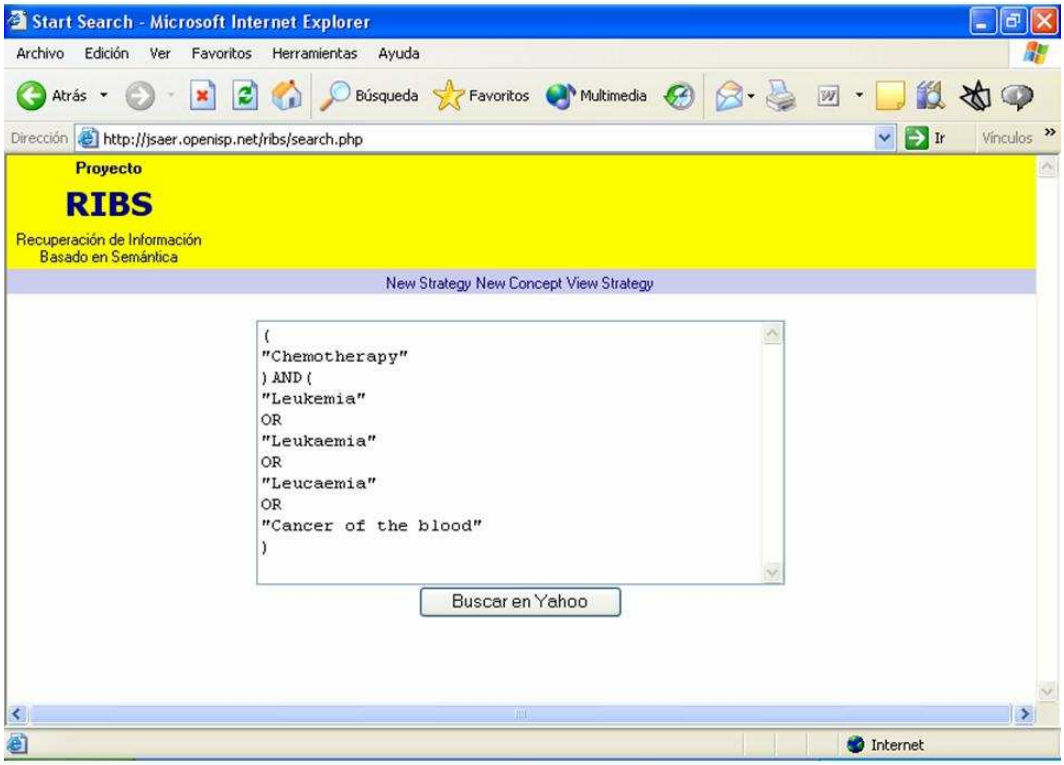


Figura A.3.6: Estrategia resultante de la consulta "quimioterapia utilizada en leucemia"

Para este prototipo, se utiliza actualmente el buscador Yahoo!, pero esta estrategia podría escribirse según las sintaxis de consulta de distintos buscadores o bases de datos donde se desea enviar la consulta.

La estrategia enviada por el prototipo al buscador corresponde al string mostrado en la Figura A.3.7.

(“Chemotherapy”) AND (“Leukemia” OR “Leukaemia” OR “Leucaemia” OR “Cancer of the blood”)

Figura A.3.7: Consulta enviada al buscador Yahoo! para el ejemplo “quimioterapia utilizada en leucemia”

Cuando se selecciona el botón *Buscar en Yahoo* el sistema muestra los resultados obtenidos tal como se observa en la Figura A.3.8.

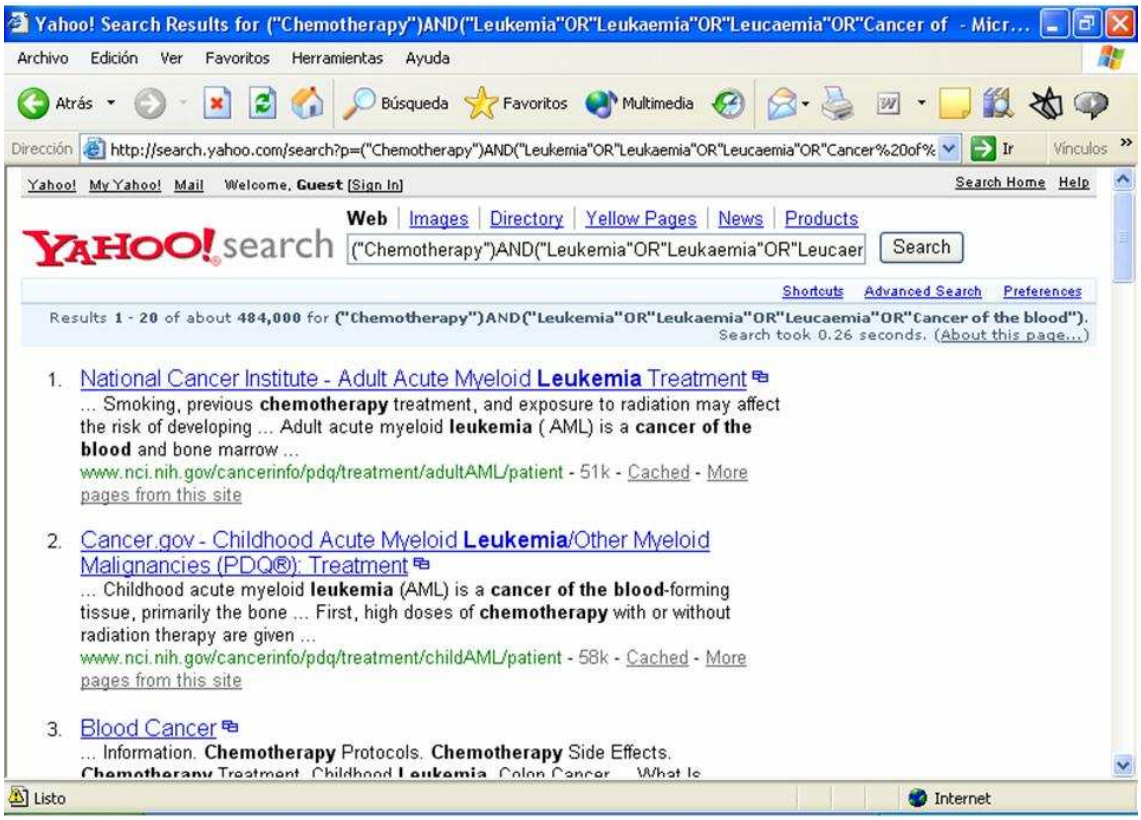


Figura A.3.8: Resultados de la consulta enviada al buscador Yahoo! para el ejemplo “quimioterapia utilizada en leucemia”