PEDECIBA Informática
Instituto de Computación – Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay

Tesis de Maestría
en Informática

Propuesta de un refinador semántico para recuperación de la información desde la Web

Claudia Deco

2004

Tutor: Regina Motz

Propuesta de un refinador semántico para recuperación de información desde la Web
Claudia Deco.

ISSN 0797-6410
Tesis de Maestría en Informática
Reporte Técnico RT 05-06
PEDECIBA
Instituto de Computación – Facultad de Ingeniería
Universidad de la República.
Montevideo, Uruguay, 2004

Tesis de Maestría

Propuesta de un Refinador Semántico para Recuperación de Información desde la Web

Claudia Deco
deco@fceia.unr.edu.ar
2004

Tutor: Dra. Ing. Regina Motz

Instituto de Computación
Facultad de Ingeniería
Universidad de la República
Montevideo - Uruguay

Abstract

As the web has become one of the biggest repository of knowledge easily accessible for everyone, Information Retrieval has stopped to be an exclusive field of information sciences specialists, and it has become a field related with any person. An information sciences specialist is a person who is in charge of translating a user's need of information into a search strategy. Maximizing the amount of obtained relevant documents for a query depends on the specialist's dexterity to prepare the search strategy. Although users do not have to know techniques of information retrieval, the proposal of this work is to improve query results by using a "specialist" that implements these techniques.

In this work, a semantic refiner is proposed. This refiner acts as a specialist in information sciences and prepares an appropriate search strategy. The semantic refiner uses linguistic resources for the preparation of this search strategy that represents the user's information need, and an improvement in the web information retrieval is achieved.

The proposed semantic refinement consists on three steps. First, it guides the user for sense disambiguation of the concepts 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 in order to increase the amount of documents to be retrieved. The linguistic resources that can be used are thesauruses, dictionaries, multilingual dictionaries and ontologies. What resources can be used, it depends on the area of knowledge of the query and on available resources for that area.

Inside this proposal, a prototype is implemented for the semantic refiner, using WordNet as the linguistic resource, and experimental results are analyzed.

Keywords: web information retrieval, query expansion, semantic refinement

Resumen

En los últimos años, 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 especialista en ciencias de la información es el encargado de expresar la necesidad de información del usuario mediante una estrategia de búsqueda. El maximizar la cantidad de documentos relevantes obtenidos para una consulta depende de la destreza de este especialista para preparar la estrategia de búsqueda. Si bien los usuarios no tienen por qué conocer técnicas de recuperación de información, la propuesta de este trabajo es la de mejorar los resultados de su búsqueda por medio de un "especialista" que implementa estas técnicas.

En este trabajo se estudia la recuperación de información en la web y se propone un refinador semántico que actúa como lo haría el especialista en ciencias de la información preparando una estrategia adecuada. Este refinador utiliza recursos lingüísticos para la preparación de esta estrategia que represente la necesidad de información del usuario, y así lograr una mejora en la recuperación de información de la web.

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.

Dentro de esta propuesta, se implementa un prototipo para el refinador semántico, eligiendo WordNet como recurso lingüístico y se analizan los resultados obtenidos para este caso.

Agradecimientos

Quisiera agradecer a mi tutora, la Profesora Regina Motz, 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 mis compañeros del grupo de investigación de la universidad, Jorge Saer y Cristina Bender.

También quisiera agradecer a la Red Iberoamericana de Tecnologías de Software para la década del 2000 (RITOS2) por el financiamiento parcial de los viajes realizados para la concreción de esta maestría.

Indice

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 ó 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, disjunción 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 ó 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 ésto 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.

Es decir, 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.

Específicamente para la recuperación de información, una posibilidad es utilizar la web y a través de ella consultar bases de datos bibliográficas ó buscadores. El usuario puede ingresar a una base de datos bibliográfica, y allí realizar 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 ó más buscadores ó 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 ó no.

Estas búsquedas, realizadas en diferentes bases de datos, tienen el problema de que cada una utiliza diferentes interfaces y diferentes índices. Por esto, el usuario deberá conocer las distintas sintaxis necesarias para cada una de las fuentes de información que consultará. Otro problema es que en el resultado de la consulta el usuario se encontrará con gran número de documentos duplicados. Por otro lado, el uso de buscadores para

localizar los documentos de interés tiene, además del problema de redundancia, el problema de necesitar descartar manualmente los documentos no relevantes a 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, ó 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, la calidad, la redundancia, la falta de estructura, la volatilidad y la heterogeneidad semántica y/o estructural. 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 de 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 [Baeza, 1998].

Si bien los usuarios no tienen por qué conocer técnicas de recuperación y extracción de información, se mejorarían los resultados de su búsqueda si por medio de una interfase que implemente estas técnicas se le sugiriera expandir ó restringir los conceptos semántica y multilingualmente y así lograr que en su respuesta los documentos recuperados sean los documentos relevantes. Esto incrementa la cantidad de documentos a recuperar. Además, se propone mejorar la precisión a través de una interacción mínima del usuario y no a través de la automatización completa de la búsqueda. En este trabajo, en particular, se realiza una experimentación para ver qué resultados arroja la utilización de WordNet[1] para ayudar al usuario en la preparación de la estrategia de búsqueda adecuada a su necesidad.

1.2. Ejemplo motivador

Supongamos que un usuario 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, ó buscadores.

El usuario ingresará a una base de datos bibliográfica, como por ejemplo Medline[2], y allí realizará su búsqueda, obteniendo una primera lista de referencias bibliográficas. Para ampliar esta búsqueda accederá a otra base de datos, como por

[1] www.cogsci.princeton.edu/~wn/
[2] Biblioteca Nacional de Medicina de Estados Unidos; www.nlm.nih.gov/medlineplus/

ejemplo Excerpta Medica[3], 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 ó más buscadores ó 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 ó no.

Todo este proceso representa varios problemas al usuario: éste deberá conocer las distintas sintaxis necesarias para cada una de las fuentes de información que consultará; tendrá que descartar los resultados no relevantes a su búsqueda; se encontrará con documentos duplicados; 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 estudiar la recuperación de información en la web y analizar la utilización de recursos lingüísticos para la preparación de una estrategia de búsqueda mediante el refinamiento semántico de los conceptos. El objetivo específico es evaluar el desempeño del recurso lingüístico WordNet para este refinamiento semántico.

El trabajo se enmarca dentro del Proyecto de Investigación Institucional “Desarrollo de nuevas metodologías para ampliar las capacidades de recuperación y extracción de información de la web” [Deco et al., 2003]. La Figura 1.1 ilustra el abordaje general del proyecto. En este contexto de trabajo, esta tesis se focaliza en uno de los módulos de esta arquitectura: el *refinador semántico*, el cual es un asistente que analiza los términos ingresados y construye una estrategia de búsqueda que represente la necesidad de información del usuario.

[3] Editorial Elsevier; www.excerptamedica.com

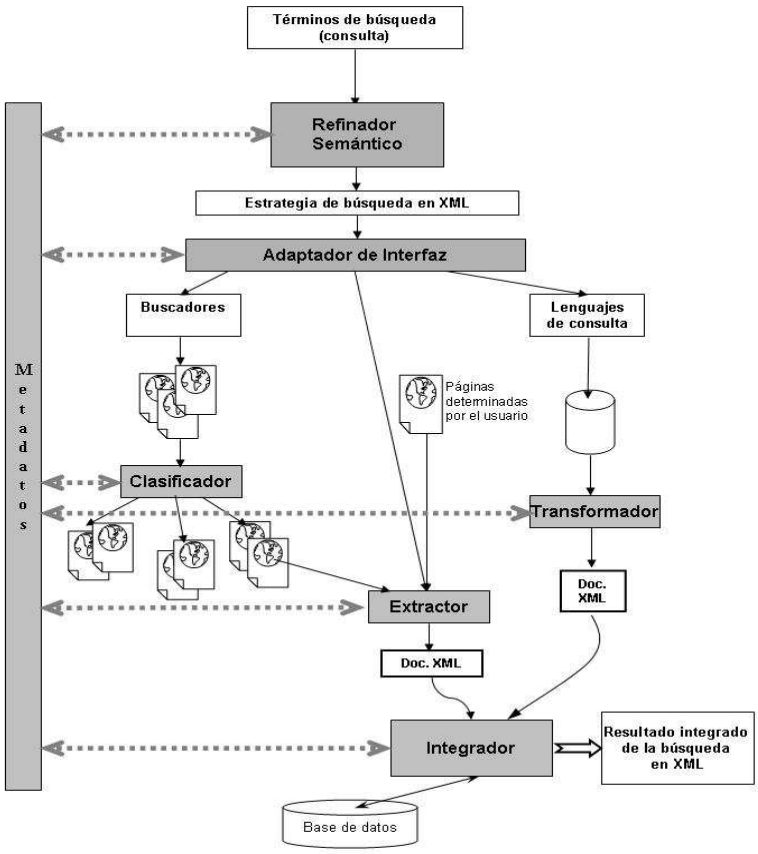


Figura 1.1: Abordaje general del proyecto [Deco et al., 2003].

El proyecto [Deco et al., 2003] propone 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, disjunción 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 Languaje) 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. Los recursos lingüísticos que se pueden utilizar son diccionarios, tesauros y ontologías. Estos recursos se tratan en detalle en el Capítulo 2.

El módulo *refinador semántico* utiliza los recursos lingüísticos para la preparación de la estrategia de búsqueda. Qué recurso ó 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 [Deco et al., 2003] los distintos tipos de fuentes que se consideran son bases de datos y páginas web. Las páginas web pueden corresponder a los resultados de una consulta realizada a través de un buscador, ó pueden ser un sitio determinado, y conocido por el usuario, de su área de interés. 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 las bases de datos disponen de lenguajes de consulta mientras que las páginas web no tienen lenguaje de consulta de su contenido.

En la consulta a bases de datos, si los resultados no están en XML se utiliza el módulo *transformador* para convertir estos resultados obtenidos a XML. En el caso de páginas web obtenidas a través de 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*. El clasificador es el encargado de agrupar los documentos resultantes según la información que contengan sobre el tema y ciertos criterios predeterminados: bibliografía, proyectos de investigación, páginas comerciales, etc. Esto evita al usuario interesado en obtener bibliografía, la revisión, por ejemplo, de páginas comerciales [Motz1 et al., 2003]. La salida de este clasificador es un conjunto de páginas web por cada categoría de clasificación. Para las páginas web correspondientes a la categoría de interés para el usuario obtenida luego de la clasificación ó para las páginas web determinadas por el usuario, un módulo *extractor* se encarga de dar estructura en formato XML a la información contenida en ellas que responda al interés del usuario. La extracción de la información deseada de páginas HTML se realiza a través de Extractores o Wrappers [Gruser et al., 1998].

Finalmente, todos los documentos XML producidos por los módulos transformador y 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 módulo refinador semántico, el cual es un asistente que analiza los términos ingresados y construye una estrategia de búsqueda que represente la necesidad de información del 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 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. 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. La solución propuesta es utilizar recursos tales como los diccionarios u ontologías, donde se pueda decidir dentro de qué contexto se está buscando el concepto ingresado por el usuario.

El objetivo de la *selección de conceptos jerárquicamente 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 para refocalizar su búsqueda y así aumentar la precisión en la recuperación.

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, utilizando recursos lingüísticos específicos del área del conocimiento disponibles en línea. 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 también puede hacerse desde el punto de vista multilingual, utilizando diccionarios multilinguales disponibles en línea para obtener dichos conceptos en otros idiomas de interés para el usuario.

Para lograr esto el refinador semántico se basa en recursos lingüísticos tales como tesauros, diccionarios, diccionarios multilinguales y ontologías para la preparación de 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.

1.4. Guía de lectura de la tesis

El resto de la tesis está organizada de la siguiente forma: en el Capítulo 2 se definen algunos conceptos básicos de Recuperación de Información y los recursos que se utilizan para preparar la estrategia de búsqueda para una consulta, tales como diccionarios, diccionarios multilinguales, tesauros y ontologías. En el Capítulo 3 se presentan trabajos relacionados. En el Capítulo 4 se propone una arquitectura para el Refinador Semántico, que permite armar la estrategia de búsqueda ampliando cada concepto semánticamente; se presenta un prototipo y experiencias realizadas con WordNet. Finalmente, en el Capítulo 5 se presentan las conclusiones y trabajos futuros. Se adjuntan a este documento dos apéndices que incluyen relevamientos de dicicionarios multilinguales y tesauros, y detalles del prototipo del Refinador Semántico.

Capitulo 2: Definiciones básicas

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 clave. En general, este proceso hacia la recuperación de documentos textuales 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 ó 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 es 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 los documentos de importancia de acuerdo con la información aportada por el usuario, como pueden ser palabras clave. 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. El usuario de este sistema puede desear recuperar un documento concreto ó un conjunto de éstos. Los documentos deseados se suelen describir mediante un conjunto de palabras clave; por ejemplo, se puede utilizar la palabra clave “cáncer de pulmón” para buscar información sobre el tema. Los documentos tienen un conjunto de palabras clave asociado y se recuperan aquellos cuyas palabras clave contengan las proporcionadas por el usuario.

La meta principal de un sistema de IR es recuperar información que podría ser útil ó importante al usuario, y no sólo datos que satisfagan una consulta dada.

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 ó en una expresión del álgebra relacional. Por ejemplo, si se consulta por la palabra “cáncer” 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 a recuperar, mientras que un sistema de recuperación de *información* recupera datos importantes que hagan la mejor coincidencia parcial con el patrón dado. Esto se debe a que la recuperación de información generalmente trata con texto de 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”, además de obtener como resultado los documentos que contengan este término, se debería obtener también los documentos en que aparezca ”neoplasma”, ”carcinoma”, 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 recuperación es obtener todos y sólo los documentos relevantes a la consulta. El problema central se reduce a establecer

una correspondencia entre el lenguaje de la consulta y el lenguaje del documento.

El usuario puede solicitar los documentos que contengan las palabras “cáncer” y “quimioterapia”, o los que contengan “cáncer” o “quimioterapia”, o los que contengan la palabra “cáncer” pero no “quimioterapia”.

La IR es una tarea compleja porque se enfrenta con varios problemas. Por un lado, los autores y los usuarios frecuentemente utilizan diferentes palabras ó 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 sobre “quimioterapia” y “cáncer” puede sustituirse por “quimioterapia” y (“cáncer” ó “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 ó a una constelación de estrellas en astronomía. Esto se soluciona desambiguando el término. Esta 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”. Otra forma de desambiguar es utilizando tesauros, cuya descripción se encuentra en el punto 2.4 de este capítulo. 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 IR 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. Esta interpretación del contenido de un documento involucra extraer información sintáctica y semántica del texto del documento y usar esta información para compararla con la necesidad de información del usuario. La dificultad radica en no sólo saber cómo extraer esta información sino también en saber cómo usarla para decidir su relevancia. Por lo tanto, la noción de relevancia es el centro de la IR. 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 IR que hasta ahora era sólo usada por bibliotecarios y especialistas en ciencias de la información. Por lo tanto, en la última década, la IR ha dejado de ser un campo exclusivo de los 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 agrupan un conjunto de páginas, las indexan, y buscan sobre éstas usando algoritmos de ranking, y muestran los documentos resultantes.

Una página web corresponde a un documento en la IR tradicional. La IR 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 ó por ser privadas.

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

Los principales problemas o 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 se puede considerar como 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.

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

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 adecuados.

Respecto a los problemas encarados por los usuarios, se pueden solucionar con el refinador semántico propuesto en esta tesis, que actúa como un “especialista en ciencias de la información”.

Algunos motores de búsqueda están potenciados por técnicas de IR, pero cubren sólo un 25% a un 55% de la web y la mayoría está en inglés [Tsikrika, 2001]. Algunos ejemplos de motores de búsqueda son: AltaVista (www.altavista.com), Excite

(www.excite.com), Google (www.google.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 índices de acceso a éstos. Los índices que se utilizan son una variante de los archivos invertidos. Al igual que en la IR se quitan las palabras no significativas ó stopwords del diccionario; se guarda la posición de las palabras en el texto, para búsquedas por adyacencia; se da un peso a los documentos para un posterior ranking de importancia en función de la cantidad de ocurrencias del término buscado ó en el lugar de la página donde aparece (en un título por ejemplo) ó en la forma en la que aparece (enfatizado o no); ó se quita peso a los documentos si su URL es muy larga lo que implicaría que es una página de poca importancia.

La búsqueda se puede realizar por una ó más palabras, raíces de palabras ó frases, que deben estar en las páginas recuperadas 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.

La información de los enlaces también se tiene en cuenta, pues un enlace representa una relación entre páginas conectadas. La principal diferencia entre los algoritmos de la IR tradicional y de la IR en la web es la presencia masiva de estos links. Estos enlaces se utilizan para dar peso a una página. En la IR 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]. Una página tiene un peso alto si la suma de los pesos de sus enlaces entrantes (ó 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 ó pocos in-links con mucho ranking.

Otro algoritmo para resolver el problema de consultas sobre un tema muy amplio que dan origen a muchas páginas recuperadas es el algoritmo llamado HITS (Hyperlink Induced Topic Search), [Kleinberg, 1998] [Kleinberg, 1999] que resuelve el problema de abundancia de documentos dando una medida de la calidad de estos, 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 almacenadas en la base de datos previo a la realización de consultas. En 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 ó 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. 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 respuesta. Esto indicaría que la gran mayoría de los usuarios desconoce las técnicas de IR, y tiene dificultad de expresar claramente su necesidad de información, y por lo tanto, no obtienen los resultados deseados.

El 85% de los usuarios de Internet utiliza motores de búsqueda para encontrar información específica [Kobashayi et al., 2000]. El mismo estudio muestra que los usuarios no están conformes con la performance brindada por los motores de búsqueda ni con la calidad de los resultados obtenidos.

Si bien los usuarios no tienen porqué conocer las técnicas de IR, 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 un refinador semántico que actúa como “especialista de ciencias de la información”. Este refinador le sugiere al usuario la desambiguación del concepto, le permite la selección de un término jerárquicamente relacionado más cercano a su necesidad, y realiza la expansión semántica y multilingual de los conceptos, a fin de preparar una estrategia de búsqueda adecuada a su necesidad de información.

Indicadores

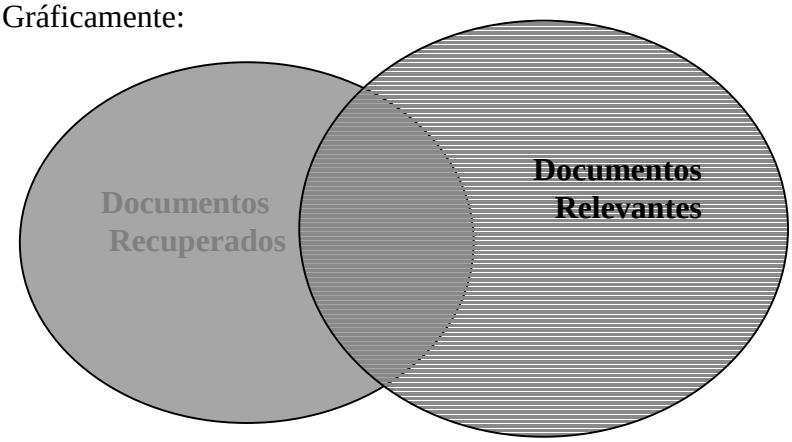
En la Recuperación de Información se han propuesto diferentes indicadores para medir cuantitativamente la performance de los sistemas de recuperación de información clásicos [Losee, 1998], la mayoría de los cuales pueden ser extendidos para evaluar los motores de búsqueda en la web. Las medidas que se definen en un modelo básico de IR 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:

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

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

Gráficamente:



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

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%

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 probar 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 debe tratar de maximizar la recuperación de documentos relevantes, y minimizar la cantidad de los documentos irrelevantes recuperados.

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.

2.3. Extracción de Información

La extracción de información o Information Extraction (IE) es una metodología que extrae información pertinente a las necesidades del usuario de grandes volúmenes de textos. Según [Bear et al., 1998] en vistas a mejorar la recuperación de información se plantea la posibilidad de la utilización de la extracción de información como un post-filtro aplicado a la salida de un sistema de IR.

El fin de un sistema de IR es consultar entre los documentos de una base de datos de gran tamaño y devolver un subconjunto de estos documentos, ordenado en forma decreciente por su relevancia respecto al tópico planteado. 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 IE es consultar un grupo de documentos, normalmente más pequeño que aquel involucrado en la búsqueda de un sistema IR, y extraer items 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 ó pueden almacenarse en una base de datos.

Mientras la IR encuentra documentos y los presenta al usuario, la IE 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 IR 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 ó 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 IE podría, con una aplicación apropiadamente configurada, completar automáticamente la planilla.

Es decir, la IR recupera documentos relevantes de las colecciones; mientras que la IE 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 IE y la IR también difieren en las técnicas que usualmente aplican. Estas diferencias se apoyan en sus objetivos y en que la mayoría del trabajo en IE 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 ha influenciado en la IR. Un factor importante para el desarrollo de la IE es el crecimiento exponencial en la cantidad de datos textuales online.

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 de interés, se necesita representar el dominio, la estructura del sitio y de las páginas web así como la terminología sobre el 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 ó textuales. Para esto se utilizan procedimientos de pattern-matching ó 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 ó listas, que permitan interpretar los datos.

Como ya se dijo, el objetivo de la IE es transformar texto sin estructura a un formato estructurado. [Eikvil, 1999] diferencia el vínculo entre la IE 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.; ó 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 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, no son gramaticales, es decir, no poseen 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 IE es esencial para cualquier esfuerzo que pretenda utilizar la web como un recurso para el descubrimiento de conocimiento. Un sistema de IE 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 IE es un conjunto de reglas de extracción

del texto ó patrones de extracción que identifican la información relevante a extraer [Soderland, 1999].

Según [Cunningham, 1999], hay cinco tareas que debe realizar la IE 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 IE, y la facilidad con la cual puede ser desarrollada, varía según el tipo de texto, el dominio ó amplitud temática del texto, el estilo en que fueron escritos los textos, por ejemplo informal ó formal, y el escenario, es decir, tipos de eventos particulares en los que el usuario de IE está interesado. Así, una aplicación particular de IE 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).

2.4. Sobre Diccionarios, Tesauros y Ontologías

En esta sección se describen recursos linguísticos tales como diccionarios, diccionarios multilinguales, tesauros y ontologías. Estos recursos se utilizan en el refinamiento semántico de los conceptos para desambiguarlos en el caso de tener varias acepciones, permitir la selección de conceptos jerárquicamente relacionados y 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” que es una parte 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 de “cáncer”.

La *hiperonimia* es una relación de superordinación entre términos, es decir un término es un hiperónimo de otro término si su significado incluye al del segundo. Por ejemplo, el término “tumor” 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 disponilbes 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 por los indizadores, que son las personas que asignan las palabras claves 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 ó 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 la calificación de homónimos para diferenciar su significado, eligiendo un significado preferido para cada término.

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, es decir grupos de palabras que representan el mismo concepto, por ejemplo agua y H2O; y homónimos, que son palabras que representan más de un concepto, por ejemplo banco, que puede referirse al mueble ó a la institución financiera.

El control de vocabulario implica la selección de un término preferido, también conocido como descriptor ó palabra clave, entre un grupo de sinónimos; y la

calificación de homónimos para diferenciar su significado, eligiendo un significado preferido para cada término.

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 adicción a drogas, la frase *drug addiction* no está permitida porque no es una palabra clave, y el tesauro indica que se debe utilizar la frase "substance dependence". Como se ve en el ejemplo, 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 por representar 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 utilizan las siglas BT (Broader Term) para indicar conceptos más amplios. 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.

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

La sigla RT se utiliza para mostrar conceptos relacionados conceptualmente con carácter horizontal. Se establecen 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 las bases de datos documentales se utilizan palabras claves para describir el contenido de un documento. Estas palabras claves, ó descriptores, pueden estar formadas por un término ó por una frase que se eligen de un diccionario de términos controlados ó permitidos para el sistema, es decir, de un tesauro. Así, el tesauro representa una herramienta documental que permite la conversión del lenguaje de un documento al propio lenguaje documentario controlado.

Los términos del tesauro se clasifican en *descriptores*, ó términos principales o preferentes o permitidos; y *no descriptores*, es decir, términos equivalentes de carácter secundario o no preferentes o prohibidos. Los términos no descriptores no pueden ser utilizados como palabras claves en los documentos para la indización del documento ni como términos de búsqueda. Para cada término no descriptor, el tesauro indica cual es el término permitido correspondiente para representar el 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 on-line. En el Apéndice 1 se presentan, agrupados según el área del conocimiento, algunos de los tesauros en línea como resultado del relevamiento realizado, indicando 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, sus definiciones y axiomas que los relacionan con otros términos que están organizados en una taxonomía y permiten realizar búsquedas con inferencias.

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 mediante palabras clave que aparecen en el código html de las páginas web dispersas en Internet. Existe, sin embargo, la tendencia de implementar el uso de metadatos para agregar datos sobre los datos; y 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 de información seguirán un esquema común, y los buscadores web compartirán con las anotaciones web los mismos esquemas.

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 el cálculo de una función que considera varios elementos de la ontología. Por ejemplo, pueden aparecer funciones como categorizar-clase, asignarfecha, etc.

Instancias: se utilizan para representar objetos 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.

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[4], OntoEdit[5], y WebOnto[6] para realizar anotaciones en documentos con lenguajes de marcado propios. Un lenguaje con gran capacidad expresiva que está emergiendo 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 lenguage de ontologías para la web, desarrollado por el W3C .

Para potenciar el uso de ontologías en la web, se necesitan aplicaciones específicas de búsqueda de ontologías, como OntoAgent[7] 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.

2.5. Utilización de los recursos

En la sección 2.4 se han descrito recursos lingüisticos 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 y páginas web, cada una de ellas con características propias.

En las bases de datos, los documentos son recopilados y analizados por instituciones especializadas que asignan las palabras claves ó términos controlados a los documentos utilizando un tesauro. En el caso de 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 término preferido que se utiliza como descriptor. La utilización de términos preferidos aumenta la precisión en la búsqueda, sin embargo si se desea aumentar la cantidad de documentos a recuperar pueden utilizarse también 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 aumentar la precisión en la recuperación de información.

Cuando la información disponible en la web no proviene de bases de datos, 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

[4] http://protege.semanticweb.org
[5] http://ontoserver.aifb.unikarlsruhe.de/ontoedit/
[6] http://kmi.open.ac.uk/projects/webonto/
[7] http://delicias.dia.fi.upm.es/OntoAgent

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. Sin embargo, 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, para definir los conceptos y sus relaciones jerárquicas. 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.

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* no proveniente de bases de datos ya que en este caso la terminología no está controlada.

Los recursos lingüísticos no sólo pueden utilizarse para la preparación de la estrategia de búsqueda como se ha explicado en la sección anterior, 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 [Motz1 et al., 2003].

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 ó 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 ó 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.

Este trabajo se focaliza en el uso de estos recursos para la preparación de la estrategia de búsqueda. 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.

Capítulo 3: Trabajos relacionados

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

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[8] es la desambiguación automática del sentido de una palabra, en la Recuperación de Información, WordNet es utilizado para expandir la query.

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] concluye 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 este autor, 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 existen términos que están relacionados, como ser “stochastic” y “statistic”, pero debido a que pertenecen a distintos grupos de Wordnet (el primero es adjetivo y el segundo es sustantivo) no se pueden relacionar.

[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[9] 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[10] es similar al recurso WordNet pero incluye vocabulario en inglés, español, alemán e italiano; también requiere licencia de uso.

[Gonzalo et al, 1998] realiza 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, demuestra que si se puede desambiguar la consulta mediante los synsets de WordNet se mejora la performance aunque no se desambiguen los documentos. Además, muestra que no se degrada la performance, si hay menos de 10 por ciento de errores en la desambiguación de sentido de la palabra.

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

[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 de sentido de una palabra basado en reconocimiento de patrones estructurado, y usan éste 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 query. 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.

[Carpineto et al, 2002] 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 de un perfil de usuario.

[Magnini y Cavaglia, 2000] presentan un trabajo cuya hipótesis es que, introducir expansiones léxicas debería traer una mejora en la recuperación de documentos relevantes. La modalidad de expansión es primero expandir cada palabra clave en 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.

CiteSeer

Algunas publicaciones de investigación en la web están disponibles en formato html, lo que permite que el texto de estas publicaciones sea buscable con los motores de búsqueda de la web. A su vez, la mayoría de los documentos de investigación publicados en la web están en formato postscript y pdf, y no en html. Por lo tanto el texto de estos documentos no es indexable por los motores de búsqueda.

CiteSeer [Bollacker et al.1998], formalmente llamado ResearchIndex, es un agente asistente para ayudar 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.

Automatiza el proceso tedioso, repetitivo y lento de encontrar y recuperar publicaciones en la web y 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 claves temáticas amplias, usa motores de búsqueda web para localizar y descargar documentos potencialmente relevantes a los temas de los usuarios.

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 clave o usar links 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 los documentos en formatos postscript ó pdf. Esta búsqueda sobre estos formatos de documentos no son realizadas por muchos motores de búsqueda y agentes. No requiere ningún esfuerzo extra por parte de los autores para la colocación de sus trabajos en la web para que sus documentos sean ingresados en esta base de datos.

La arquitectura del agente tiene tres componentes principales: Un subagente para localizar y adquirir automáticamente publicaciones de investigación. Un parser de documentos y creador de la base de datos. Y una interface de navegación para la base de datos que soporta la búsqueda por clave y la navegación por links de citación.

Cuando el usuario desea explorar en 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 ó 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 ó pdf. Luego, verifica que este texto ASCII sea un documento de investigación, incluyendo un chequeo de la

existencia de referencias ó 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 ó 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 ó sobre las citas bibliográficas. Después de una búsqueda inicial por palabras clave 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 por 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 resultar en una medida de distancia que sea más precisa que un método por si solo. Por último, hay que resaltar que la implementación de este agente está en uso y el motor de búsqueda está disponible en http://citeseer.org/.

InfoSleuth

InfoSleuth[11] [Nodine et al. 1998] [Nodine et al. 1999] [Fowler et al. 1999] es un sistema prototipo basado en agentes, diseñado para integrar fuentes y herramientas heterogéneas y distribuidas a través del uso de ontologías comunes. Un conjunto de agentes de InfoSleuth colabora en el nivel semántico para ejecutar la recolección de información y 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 metadata esquemático 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 de Knowledge Query Manipulation Languaje (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 interface genérica para sistemas de representación de conocimiento subyacente.

La arquitectura de InfoSleuth es dinámica y basada en agentes. Cada agente en InfoSleuth provee un conjunto de servicios que se pueden describir como un conjunto de tareas sobre el dominio de la interacción InfoSleuth. El agente de usuario asiste al usuario en las consultas utilizando ontologías y muestra los resultados. El agente de broker hace corresponder solicitudes de servicios o información con agentes que pueden proveerlos. El agente de ontología provee conocimiento y responde consultas sobre ontologías. Los agentes de recursos traducen las consultas y datos almacenados en algún repositor de datos externo y el repositorio propio.

El agente de consulta de recursos múltiples maneja la descomposición y distribución de subconsultas para agentes de recursos varios y luego recompone los resultados. 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 soportada denotando la intermediación semántica, mediante brokers, lo que permite a los agentes identificar potenciales colaboradores.

Ontobroker

Es un sistema de búsqueda basado en ontologías con axiomas. Ontobroker[12]

[11] www.argreenhouse.com/InfoSleuth/
[12] http://ontobroker.semanticweb.org

[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 web crawler se encarga de recolectar páginas web, extraer las descripciones semánticas y parsearlas al formato interno de Ontobroker. La 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 de 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 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 está disponible en la web y 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 definida que pueden entenderse directamente y procesarse por agentes automáticos u otras herramientas de software.

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 comenzó en el '96 justo con el comienzo de la era JAVA y se adoptó esta 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[13], 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, descripto 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 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 descriptos 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 ó 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.

WebFind

WebFind [Monge et al. 1996] es una herramienta que descubre documentos científicos que están disponibles por sus autores en la web. Es un proyecto de la

[13] http://mozart.isi.edu:8003/sensus2/

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.

WebFind utiliza la información de estas fuentes para encontrar un camino para descubrir los documentos en la web. 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 clave 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 búsqueda se ejecuta mientras el usuario espera una respuesta a su consulta, y 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 links hasta que el artículo requerido es encontrado. La búsqueda se 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 link 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 links 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 link. El contexto del link es su texto ancla, ó anchor text, y las dos lineas 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 tipo de búsqueda es que tiene una baja performance debido a que la búsqueda se realiza online.

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[14].

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, como por ejemplo un periódico personalizado, que correspondan al interés del usuario. 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. El controlador applet interactúa con el usuario. A través de este controlador, el usuario puede expresar sus intereses cuando navega y proveer 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 continuamente. 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 top en el perfil actual y lo 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

[14] http://www-2.cs.cmu.edu/~softagents/webmate.html

todas y cada una de las URLs en el resultado. Luego calcula la similaridad del perfil y recomienda las páginas con una similaridad 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.

Untangle

El proyecto Untangle[15] [Welty, 1996] [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 clave 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). Classsic 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 interface 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 de 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: navegando la web. La contribución de este trabajo es: fabricar una técnica KR aprovechable en la web puede demostrar potencialmente a muchas comunidades el beneficio práctico de usar KR&R. La ontología para información electrónica combinado con una taxonomía de temas lleva el estándar de las ontologías bibliográficas un paso más allá. Después de llevar más allá la práctica del testeo y uso, la nueva ontología será sometida a un análisis más formal y riguroso para sumisión de la librería ontológica Ontolingua.

La motivación original para mover esta investigación de la distribución por mail 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 IR. El proyecto Untangle espera ser un ejemplo de la utilización de KR&R, pero esto va a ser difícil ya que las Bibliotecas Digitales y la web en general están fuertemente ligadas a la IR, pero las capacidades de inferencia y conocimiento del KR aún tienen mucho más por ofrecer.

[15] http://untangle.cs.vassar.edu

Metabuscadores

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 interface 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.

Comparación entre los distintos proyectos

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

[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

Figura 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 y 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 concepto. Ninguno le permite al usuario precisar su interés de búsqueda seleccionando un concepto jerárquicamente relacionado.

La propuesta 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 ó nulos, sugiriendo al usuario el concepto 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 linguísticos 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 será recompensado evitándole a posteriori la lectura y la clasificación manual de los documentos que no sean de su interés.

Capítulo 4: El Refinador Semántico

4.1. Arquitectura del Refinador Semántico

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 construye 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 precisar 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 tales como los diccionarios u ontologías, donde el usuario pueda decidir dentro de qué contexto se está buscando el concepto ingresado. Esta decisión la realiza en forma interactiva el usuario.

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 la recuperación de información. Además, si el usuario desea obtener información en más de un idioma, entre estas expansiones se pueden incorporar la traducción de dichos términos. Los recursos lingüísticos utilizados aquí son diccionarios, tesauros, ontologías y diccionarios multilinguales. Esta expansión la realiza el refinador en forma automática.

Finalmente, el resultado es una estrategia de búsqueda preparada en forma automática por el refinador.

Los recursos lingüísticos que pueden utilizarse son tesauros, diccionarios, diccionarios multilinguales y ontologías. Qué recurso ó recursos utilizar, depende del área del conocimiento de la consulta y de los recursos disponibles para ese área. Por ejemplo, para una consulta sobre temas médicos, se puede utilizar el tesauro MeSH[18].

La arquitectura propuesta del refinador semántico y que refleja las tareas que debe realizar, se presenta a continuación en la Figura 4.1, donde los módulos sombreados indican que se necesita participación del usuario.

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

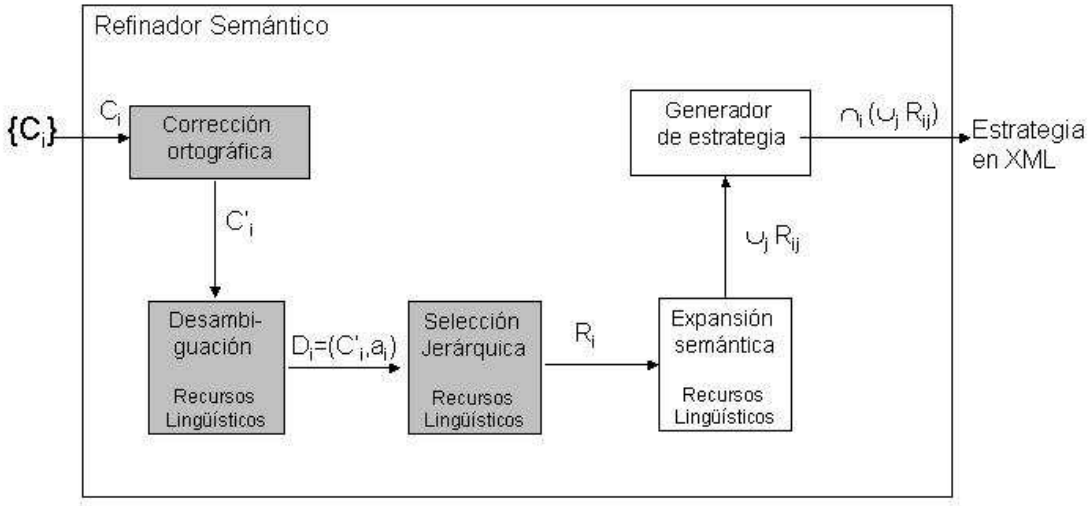


Figura 4.1: Arquitectura del Refinador Semántico para la construcción de la estrategia de búsqueda

Para realizar una consulta, el usuario ingresa un conjunto de conceptos $\{C_i\}$ con $1 \leq i \leq n$ y la salida del *Refinador Semántico* es una estrategia de búsqueda asociada a estos conceptos. Una estrategia de búsqueda es una expresión lógica compuesta por distintos conceptos combinados con los conectores lógicos de conjunción, disjunción y negación.

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.

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. El usuario selecciona la acepción que corresponde a su interés de búsqueda. Cada acepción de un concepto tiene una jerarquía conceptual asociada y 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.

Para realizar esta desambiguación se utilizan recursos lingüísticos: tesauros, diccionarios, diccionarios multilinguales y ontologías. Qué recurso ó recursos utilizar, depende del área del conocimiento de la consulta y de los recursos disponibles para ese área. Un recurso 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.

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 ó inferior, ó 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, 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, ó de un área específica del conocimiento, por ejemplo MeSH para el área salud.

Expansión semántica:

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

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 ó más recursos lingüísticos para la incorporación de estos sinónimos.

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, el Generador de estrategia 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 y C_2 y .... y (no C_h) y ... y C_n

planteada por el usuario, se obtiene la estrategia siguiente:

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

AND

.....

AND

(NOT (R_{h1} OR R_{h2} OR ... OR R_{hr}))

.....

AND

(R_{n1} OR R_{n2} OR ... OR R_{nr})

donde:

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

...

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

...

(R_{n1} OR R_{n2} 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 C1>
                <ampliación 1> R11 </ampliación 1>
                .....
                <ampliación r> R1r </ampliación r>
        </concepto C1>
        <concepto C2>
                <ampliación 1> R21 </ampliación 1>
                .....
                <ampliación r> R2r </ampliación r>
        </concepto C2>
        ......
        <no concepto Ch>
                <ampliación 1> Rh1 </ampliación 1>
                .....
                <ampliación r> Rhr </ampliación r>
        </no concepto Ch>
        ......
        <concepto Cn>
                <ampliación 1> Rn1 </ampliación 1>
                .....
                <ampliación r> Rnr </ampliación r>
        </concepto Cn>
</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 (Figura 1.1, pág. 8). Este módulo es el *Adaptador de interfaz*, y se encarga de traducir la estrategia de búsqueda a la sintaxis de las distintas fuentes.

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 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 ó 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 disjunción (OR) cuando debería haber usado conjunción (AND) ó 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 ó ningún documento como resultado de una consulta, pudo haber cometido también errores de estrategia ó 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 disjunción (OR), ó 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 “color” y “colour”), ó 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 disjunción 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.

4.2. 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 este 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” y “lung neoplasms”.

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 siguiente estrategia de búsqueda:

lung cancer OR carcinoma of the lungs

OR cáncer de pulmón OR lung neoplasms

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 }

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).

The noun "lung cancer" has 1 sense in WordNet.

1. lung cancer -- (carcinoma of the lungs; one of the commonest forms of cancer)

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 autómatica 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 ó 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

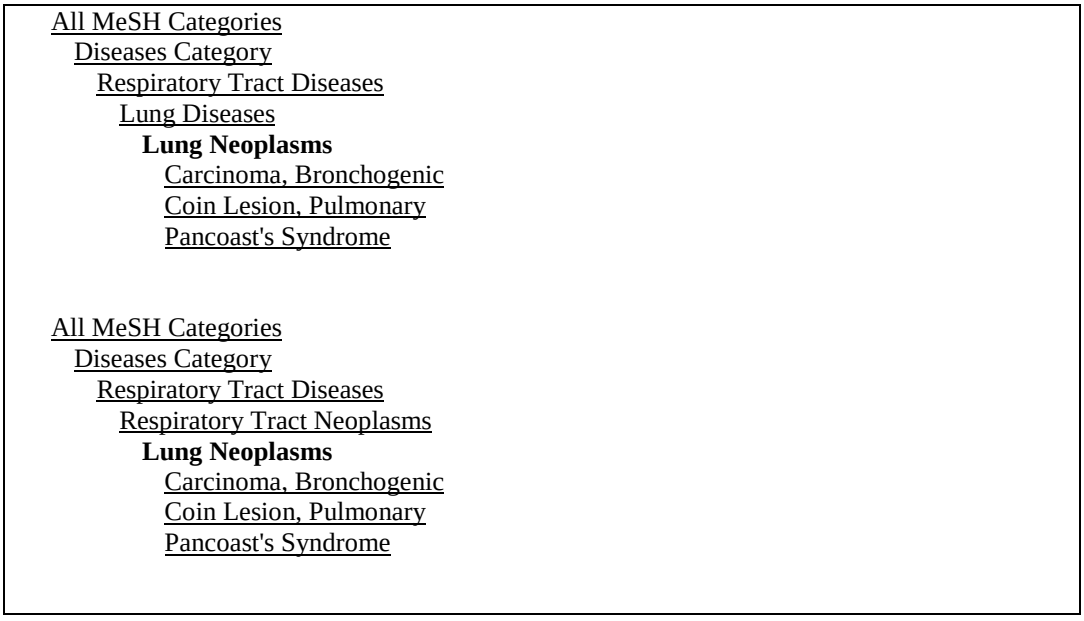


Figura 4.7: Una vista del tesauro MeSH de Medline

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 utilizando como recurso un diccionario la frase: “pulmonary blastomas” y utilizando como recurso un diccionario multilingual la frase: “blastoma pulmonar”.

Finalmente, el módulo Generador de estrategia, construye la siguiente estrategia de búsqueda:

pulmonary blastoma OR pulmonary blastomas OR blastoma pulmonar

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:

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

Ejemplo 3

Los ejemplos anteriores son sencillos pues la estrategia está formada a partir de *un solo* concepto. 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 se combinan en el módulo Generación de estrategia. Como resultado, la estrategia de búsqueda asociada consta de la disjunción 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*. Realizando un procedimiento similar al mostrado en los ejemplos anteriores por cada uno de estos conceptos, la estrategia de búsqueda resultante 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)

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

4.3. 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. Uno de los problemas presentados en este desarrollo fue que, al utilizar recursos libres y gratuitos disponibles en la web, éstos pueden dejar en algún momento de estar disponibles.

El prototipo desarrollado se describe en el Apéndice 2. Se discuten a continuación algunas de las experiencias llevadas a cabo durante el proceso de desarrollo del prototipo de refinador semántico, el cual debe cubrir las tareas explicadas antes: Corrección Ortográfica, Desambiguación del concepto, Selección de jerarquía, Expansión Semántica y Generación de Estrategia.

El primer paso fue buscar en la web qué recursos computacionales y/o de información estaban disponibles para utilizar en el prototipo. Para la correción ortográfica se utiliza el método Spelling Suggestion del web service de Google[19]. Para la desambiguación, la selección de jerarquía y la expansión semántica se utiliza el recurso lingüístico WordNet 1.6, en formato RDF.

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

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

El prototipo [Saer, 2004] se adjunta en el CD-ROM anexo, y puede ser consultado en jsaer.openisp.net/ribs1.1. 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.

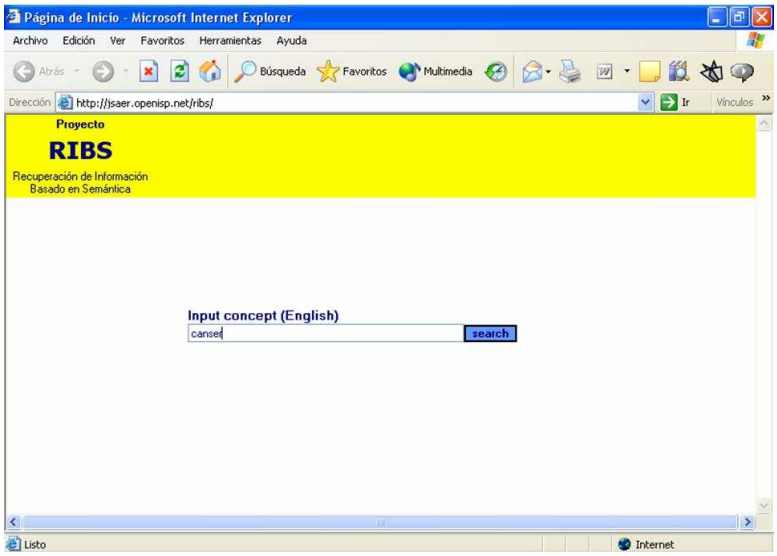


Figura 4.9. Pantalla inicial del prototipo de refinador semántico

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

[19] www.google.com/apis
[20] www.php.net

- El usuario ingresa, en inglés, el concepto a buscar.
- Con el botón de búsqueda “search”, envia 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. Más detalles se pueden consultar en el Apéndice 2.

4.4. Experimentación

El refinamiento semántico tiene por objetivo formular una estrategia de búsqueda a partir de conceptos ingresados por el usuario. En esta tesis se proponen distintos recursos lingüísticos tales como tesauros ú ontologías para este fin. Un recurso muy utilizado y con el cual se realizaron las experiencias que se describen en esta sección, es WordNet.

El refinamiento semántico resuelve muchos de los problemas que se presentan en la formulación de una estrategia de búsqueda, como ser el correcto uso de la disjunción 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 mismos.

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. El recurso WordNet incluye sinónimos, variantes de escritura de nombres propios, ampliación de siglas, variaciones de deletreo, y para ciertos términos su escritura en otros idiomas.

La precisión de los resultados se logra presentándole al usuario una estructura jerárquica de conceptos que le permite hacer un recorrido conceptual de su consulta. Es decir, moverse por jerarquías conceptuales, subiendo ó bajando de nivel conceptual, y seleccionando un término más preciso a su necesidad de información. El recurso WordNet tiene una jerarquía conceptual, y muestra para cada término sus términos específicos ó hipónimos, y su término más amplio ó hiperónimo.

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.

En la tabla siguiente (Tabla 4.1), se presentan las consultas realizadas. A continuación se presentan observaciones sobre las consultas desde el punto de vista de la cantidad de documentos recuperados y desde la cantidad de documentos relevantes en los primeros 50 documentos.

Tabla 4.1: 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
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

		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
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
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

		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
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
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

		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
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

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 ó 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 Marquez* no se encuentra en WordNet, pero el refinamiento semántico agrega a esta frase las comillas: *"Gabriel García Marquez"*. 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 ó 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 ésto 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, ó 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 polisémicos 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* ó *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 4.10. 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 4.11 se muestran estos mismos resultados, pero con escala logarítmica para no descartar y poder visualizar la consulta 1.

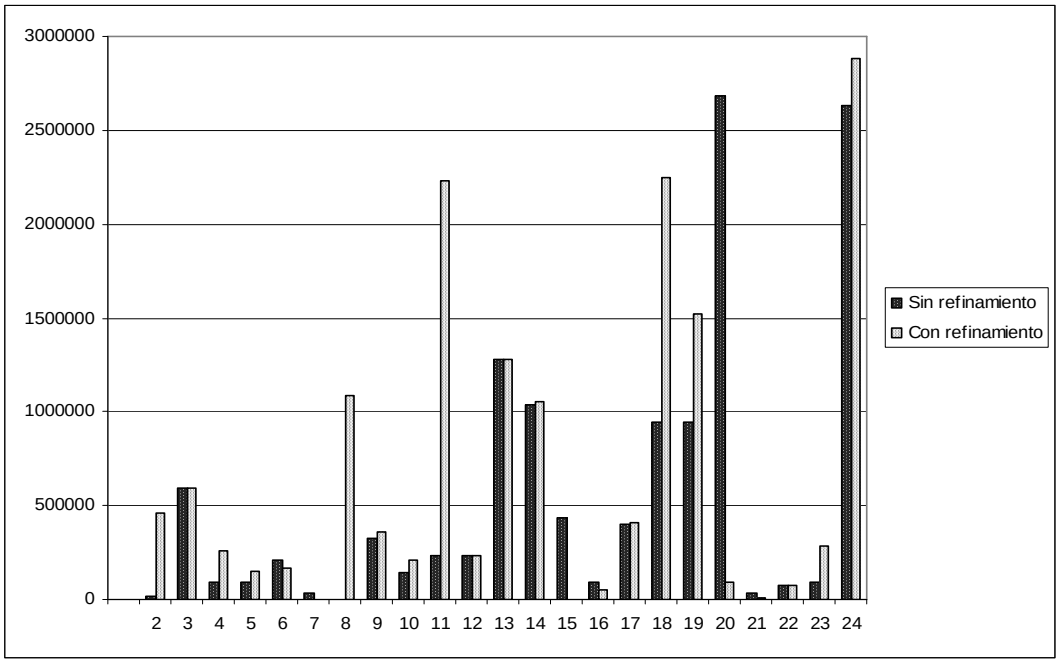


Figura 4.10: Cantidad de documentos resultantes con y sin refinamiento semántico

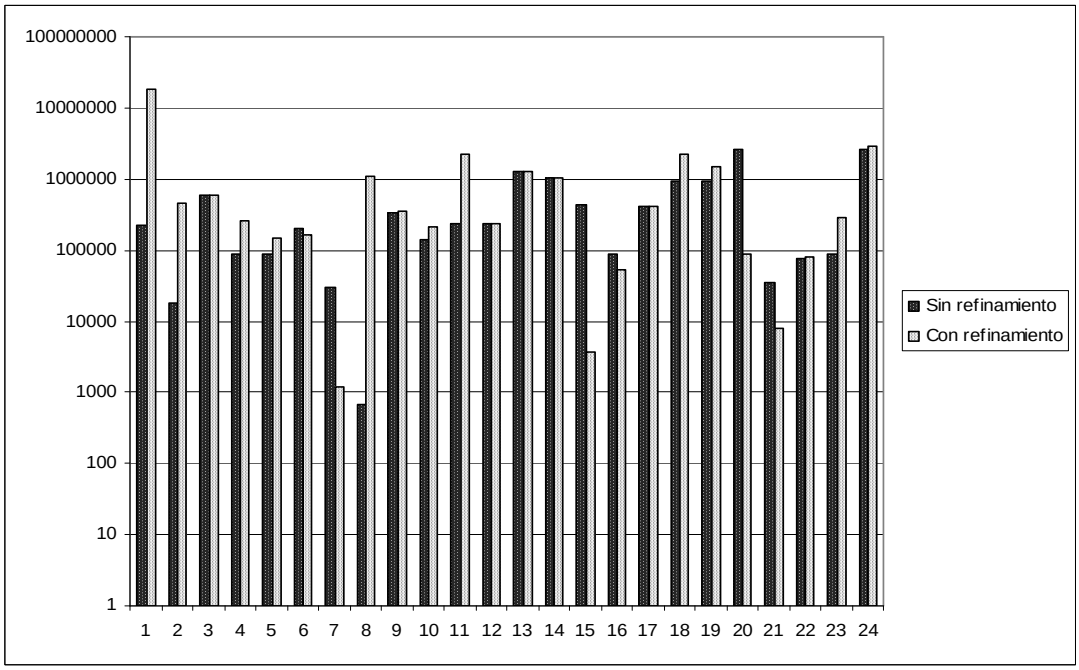


Figura 4.11: Cantidad de documentos resultantes con y sin refinamiento semántico en escala logarítmica

Cantidad de documentos relevantes recuperados en los 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 precisa.

- *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, ó 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 ó 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 ó siglas.

En la Figura 4.12. 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.

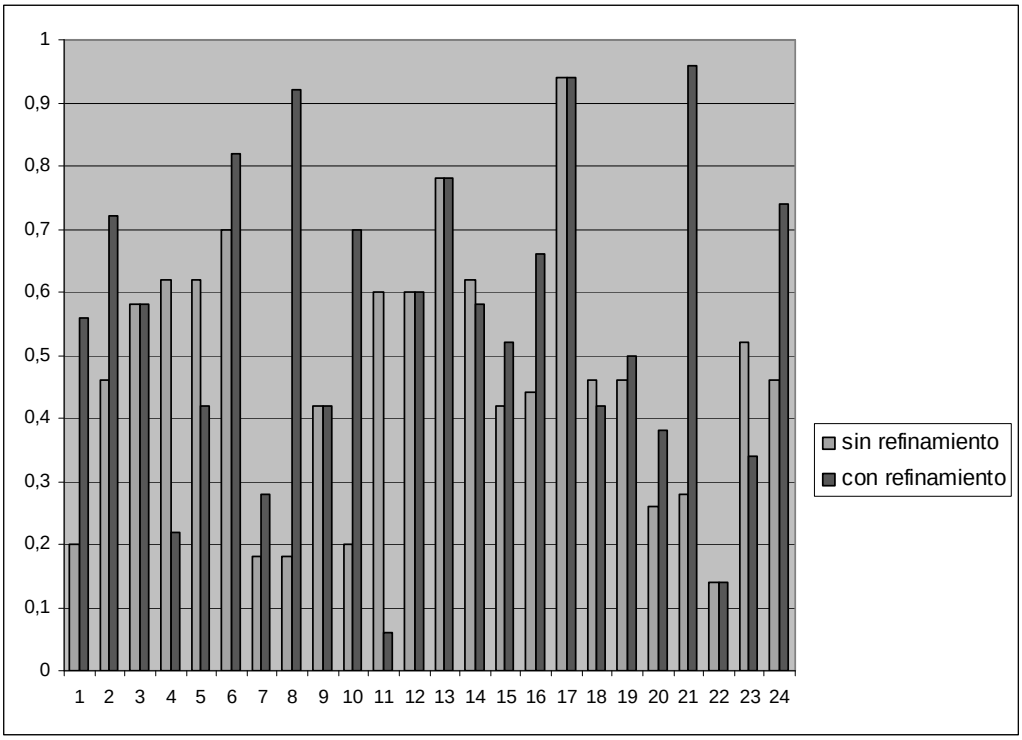


Figura 4.12: Precisión en los primeros 50 documentos,con y sin refinamiento semántico

Conclusiones de la experimentación

El objetivo de las experiencias realizadas es evaluar la utilización del recurso lingüístico WordNet para la preparación de la estrategia de búsqueda para la recuperación de información en Internet. 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 ó 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 4.2. y en la Figura 4.13.

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

Tabla 4.2: Promedio cantidad de documentos recuperados según cantidad de conceptos utilizados

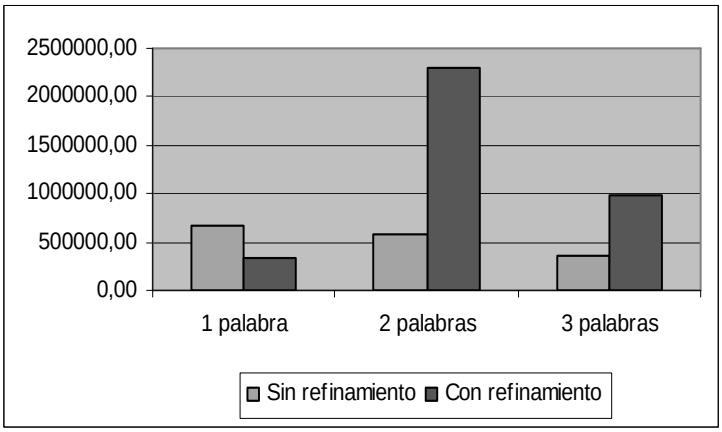


Figura 4.13: Promedio cantidad de documentos recuperados según cantidad de conceptos utilizados

- Analizado el número de conceptos utilizados en cada consulta, el promedio de la precisión aumentó luego del refinamiento semántico para consultas que involucran uno ó 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 4.3. y en la Figura 4.14.

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

Tabla 4.3: Promedio de precisión según cantidad de conceptos utilizados

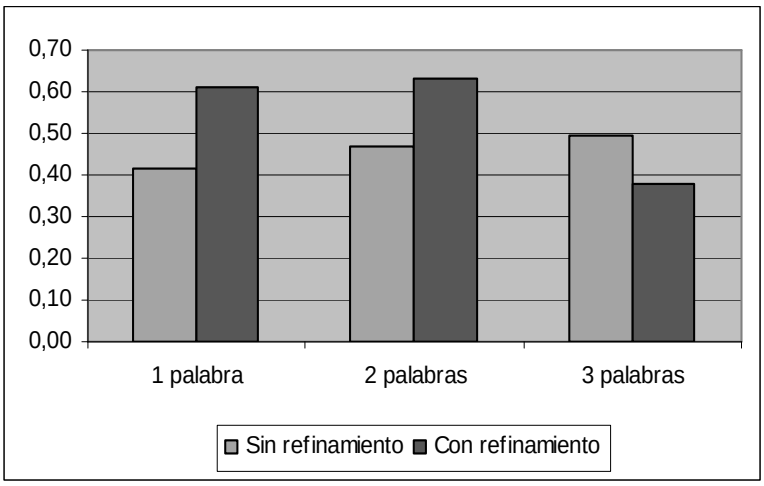


Figura 4.14: Promedio de precisión según cantidad de conceptos utilizados

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 4.4.

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 4.4: 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 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.

Capitulo 5: Conclusiones y trabajos futuros

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 especialista en ciencias de la información es el encargado de expresar la necesidad de información del usuario mediante una estrategia de búsqueda. Una búsqueda 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 de este especialista para preparar la estrategia de búsqueda.

Si bien los usuarios no tienen 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 refinador semántico propuesto es el que actúa como lo haría el especialista en ciencias de la información preparando una estrategia adecuada.

En esta tesis se experimenta el impacto de utilizar un refinador semántico basado en WordNet para mejorar los resultados de la búsqueda y se presenta un prototipo basado en este recurso como una aplicación concreta. Además, este trabajo aporta un estudio en los temas de Recuperación de Información.

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 multilinguales y ontologías. Qué recurso ó 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.

En este trabajo, se presentó la arquitectura de este refinador semántico en la Figura 4.1 (pág. 40). Este refinador semántico está dentro de una arquitectura general, presentada en la Figura 1.1. (pág. 8), que es una propuesta para mejorar la recuperación de información a partir de distintas fuentes brindándole al usuario una única interfase de consulta y obteniendo una respuesta integrada.

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. Las experiencias realizadas se presentan en el Capítulo 4. Cabe destacar que en los resultados generales tanto el promedio de la cantidad de documentos recuperados como la precisión se incrementan en cerca de un 20%. Esto muestra que la propuesta de refinamiento semántico presentada mejora la recuperación de información de la web al utilizar WordNet como recurso lingüístico para la preparación de la estrategia de búsqueda.

Sin embargo, 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 no están incluidos y WordNet carece de un área del conocimiento. 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 del conocimiento qué recurso lingüístico especializado utilizar.

Trabajos Futuros

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.*

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 si no se recupera información suficiente.

El refinador semántico resuelve la mayoría de los problemas relacionados con estas contingencias: la desambiguación de términos ambiguos o no específicos, el correcto uso de la disjunción 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.

Queda abierto el problema de que en caso de obtener como resultado pocos ó 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, ó que estén separados por una cierta cantidad de palabras uno de otro.

- *Utilización de perfiles de usuario.*

En el refinamiento semántico actualmente propuesto, no existe una identificación del usuario. 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 del usuario, son ingresados manualmente por éste. Otra posibilidad es utilizar información disponible desde una fuente de datos para extraer los conceptos iniciales a ingresar al refinador semántico.

Una aplicación posible es la búsqueda asistida de evidencia clínica en medicina, facilitando el acceso de los usuarios médicos a los contenidos de información disponibles en los repositorios de literatura científica. El esfuerzo que exigen los avances del saber médico y la fugacidad de la vigencia de los criterios científicos, conlleva la necesidad de una constante actualización de sus contenidos. La literatura científica de acceso electrónico, puede proporcionar una buena solución a este problema.

Es decir, incorporar al refinador semántico mecanismos que utilicen la información disponible de un paciente a partir de una historia clínica electrónica, en un formato tal como el pre-estándar ENV 13606 de la Unión Europea [Plüss et al., 2003], para generar una estrategia de búsqueda. Esta estrategia debería generarse automáticamente a partir de ciertos campos de la historia clínica electrónica que elija el usuario médico.

- *Expansión multilingual.*

En el refinador semántico propuesto en esta tesis, se ha considerado la expansión multilingual de los conceptos ingresados por el usuario en el módulo expansión semántica. Todo el proceso del refinamiento semántico se realiza en un solo idioma y se considera el concepto en otros idiomas como un sinónimo más. La recuperación de información multilingual en la web es un tema de investigación abierto.

Así como en un solo idioma existe el problema de determinar la acepción de interés de un término, es decir desambiguar el concepto, en la traducción a otros idiomas ocurre un problema similar. Por ejemplo, el término “investigación”, en inglés puede escribirse como “investigation” ó como “research”, con significados distintos. El primero se refiere a una investigación policial y el segundo a una investigación académica. Por lo tanto, hay que introducir una manera de desambiguar el concepto al traducirlo a otros idiomas. Una posibilidad es que el usuario decida cuál es la acepción de la traducción de su interés. Otra posibilidad sería realizar ésto en forma automática, detectando la acepción a partir de los otros conceptos ingresados por el usuario en la consulta ó a partir de perfiles de usuario.

- *Feedback de relevancia*.

Una forma de mejorar la estrategia de búsqueda es a partir de un feedback de relevancia por parte del usuario. El feedback de relevancia es una técnica de la recuperación de información clásica que reformula una consulta en base a documentos identificados por el usuario como relevantes [Salton, 1983]. El feedback de relevancia ha sido y todavía lo es un gran área de investigación activa en la recuperación de información. Es usado ampliamente y con resultados exitosos en la recuperación de información tradicional, pero aún no ocurre lo mismo en la recuperación de información de la web.

Sería de interés utilizar el feedback de relevancia para mejorar la estrategia de búsqueda producida por el refinador semántico. En este caso, el usuario indicaría un subconjunto de documentos resultantes de la búsqueda que respondieron a su interés, y se extraerían términos de este subconjunto de documentos para incorporarlos 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.

- *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 ó sinónimos, incorporar a la estrategia de búsqueda nuevos conceptos obtenidos a través de la inferencia.

- *Enfoque de agentes*.

Otra aproximación al problema es el uso de agentes inteligentes. 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 es un agente de software que puede funcionar fiablemente en un entorno rápidamente cambiante e impredecible, como es la web.

Una de las tareas que podría ser abordada mediante agentes es la exploración automática de la web para recuperar las páginas relevantes ante una necesidad de información de parte de un usuario. El usuario delega en el agente, después de haberle facilitado algunas instrucciones, como por ejemplo indicándole qué tipo de información se desea. Los agentes pueden configurarse con diferentes perfiles para tomar decisiones de acuerdo a las necesidades del usuario y hacer tareas más específicas y personalizadas.

Los agentes pueden tener un itinerario que les permite viajar en algún esquema particular de búsqueda y acceso de información, yendo a servidores en forma directa o a servidores más cercanos que permitan mayor rapidez en el acceso de la información y el manejo de los recursos. Otro tema importante es el paralelismo: el hecho de poder enviar varios agentes a varios lugares para hacer tareas específicas, permite aprovechar la potencialidad de las máquinas en paralelo.

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, 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

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

[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.

[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.

[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, Univ. 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

[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 et al., 2003] Deco, C., Bender, C., Chiari, M., Fornari, J., Saer, J. Reporte Técnico número 3 del Proyecto: Desarrollo de nuevas tecnologías para ampliar las capacidades de recuperación y extracción de la información de la web. Departamento de Investigación Institucional, Facultad de Química e Ingeniería de Rosario, Universidad Católica Argentina, Junio, 2003.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

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

[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.

[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.

[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

[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.

[Motz1 et al., 2003] 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.

[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., 1998] Nodine, m., Perry, B., Unruh, A, Experience with the InfoSleuth agent architecture. In AAAI-98 Workshop on Software Tools for Developing Agents, 1998.

[Nodine et al., 1999] Nodine, M., Fowler, J., Perry,B. Active information gathering in InfoSleuth. In Proceedings International Symposium on Cooperative Database Systems for Advanced Applications, 1999.

[OWL] OWL Web Ontology Language 1.0 Reference, http://www.w3.org/TR/2002/WD-owl-ref-20020729/ Consultado el 08-09-03.

[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.

[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.

[Saer, 2004] Saer, J. Prototipo Refinador Semántico. Desarrollado en el proyecto RIBS: Recuperación de información basada en semántica. Director del proyecto Claudia Deco. Departmento de Investigación Institucional. Facultad de Química e Ingeniería de Rosario. Universidad Católica Argentina. 2004.

[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.

[SemanticWeb] http://www.semanticweb.org/. Consultado el 08-09-03.

[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

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

[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.

[W3C 1999] http://www.w3.org/TR/1999/REC-xslt-19991116 Recomendaciones del Consorcio W3C sobre XSLT. Consultado el 11-06-03.

[W3C] Extensible Markup Language (XML), Recomendaciones del Consorcio WWW. http://www.w3.org/TR/REC-xml. Consultado el 11-06-03.

[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.

[Welty, 1996] Welty, C. Intelligent Assistance for Navigating the Web. Proceedings of The 1996 Florida AI Research Symposium. May, 1996.

Apéndice 1: Relevamiento de Diccionarios Multilinguales y Tesauros disponibles[21] en la Web

Relevamiento de Diccionarios multilinguales disponibles en la Web

- **Foreignword.com**
  Traductor que facilita la búsqueda mediante un sistema de búsqueda por palabras. Contiene varios idiomas, entre ellos el castellano. http://www.foreignword.com/
- **EuroDicAutom**
  Base de datos terminológicos de la Unión Europea, permite traducir términos en los idiomas de los países que forman la Unión Europea. http://eurodic.ip.lu/cgi-bin/edicbin/EuroDicWWW.pl
- **Histopia - Diccionario Multilingüe**
  Contiene términos en más de quince idiomas, entre ellos el castellano. http://www.histopia.nl/onldict/
- **Biblioteca Popular en Internet**
  Recopilación de diccionarios de lenguas de la Unión Europea. http://www.arrakis.es/~margaix/ficheros/diccion.htm
- **Diccionarios.com**
  Permite traducir términos del castellano al inglés y francés. Contiene un diccionario general de lengua española y de sinónimos. http://www.diccionario.com/
- **Cyberdisco, Online Word Translator**
  Traductor virtual que transforma simultáneamente a alemán, español, francés, inglés e italiano. http://signserver.univ-lyon2.fr/home/Traduc.html
- **AllWords.com**
  Traductor virtual que transforma a alemán, español, francés, holandés inglés e italiano. http://www.allwords.com/
- **Rivendell's Machine Translation Dictionary**
  Traduce cualquier término alemán, español, francés, italiano e inglés. http://rivendel.com/~ric/resources/translator.html
- **Online English to Spanish to English Dictionary**
  Diccionario multilingüe de inglés con otros idiomas, incluido el español. http://www.freedict.com/onldict/spa.html
- **Babylon, Diccionario y Traductor**
  Diccionario multilingüe. Puede usarse en línea ó puede ser descargado. http://www.babylon.com/
- **WordReference.com**
  Traductor en línea multilingüe que traduce los idiomas español, alemán, italiano y francés al inglés y viceversa. http://wordreference.com/
- **LangSoft Multilingual Dictionary**
  Traduce términos en varios idiomas: Alemán, castellano, francés, inglés, italiano y ruso. http://www.translator.cz/translator.shtml

[21] Relevamientos realizados en septiembre de 2003

- **Online dictionary, The**
  Traductor virtual que traduce simultáneamente a alemán, español, húngaro e inglés. http://dictionary.lezlisoft.com/dictionary/
- **NETGLOS**
  Traductor multilingüe de términos relacionados con la informática. Contiene varios idiomas entre los que se encuentra el castellano. http://wwli.com/translation/netglos/
- **Zaz - Dicionários**
  Traductor virtual que agrupa los idiomas alemán, castellano, francés, inglés, italiano y japonés. http://www.zaz.com.br/dics/
- **Terminology Collection: Online Dictionaries**
  Catálogo actualizado periódicamente de diccionarios generales de lengua y diccionarios especializados o terminologías.
  http://www.uwasa.fi/comm/termino/collect/
- **Proyecto ARTFL - Diccionario Francés-Inglés**
  Vocabulario constituido por 75.000 términos del francés y el inglés. Realizado por la Universidad de Chicago. http://humanities.uchicago.edu/forms_unrest/FR-ENG.html

Diccionarios multilinguales especializados:

- **Diccionario técnico textil**
  Español, inglés, francés, alemán.
- **Eurodicautom**
  Diccionario multilingüe del Consejo Europeo y del Parlamento Europeo (alemán, danés, español, finlandés, holandés, inglés, italiano, portugués y sueco)
- **EUR-Lex**
  Lenguaje legal de la Unión Europea en 11 idiomas.
- **Fishbase Glossary Searched Term**
  Inglés, francés, español, portugués.
- **ILOTERM**
  Terminología social y laboral: inglés, español, francés, italiano, alemán.
- **Inmobilienlexikon**
  Léxico inmobiliario en alemán, francés, italiano y español.
- **ITU Telecommunication Terminology Database (TERMITE)**
  Inglés, francés, español, ruso.
- **Multilingual Glossary of technical and popular medical terms in 9 European Languages**
  Danish - Dutch - English - French - German - Italian - Portuguese and Spanish.
- **TermFinance**
  Glosario de términos financieros en inglés, alemán, francés, italiano.

Relevamiento de Tesauros disponibles en la Web

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

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

- 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

http://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

http://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

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

- Tesauro de la Materia Familia y Civil

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

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: Prototipo

En este apéndice se describe el prototipo [Saer, 2004] para el refinamiento semántico de conceptos. La entrada son los conceptos ingresados por el usuario y la salida del prototipo es una estrategia de búsqueda.

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[22]. 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.

Estos servicios fueron ensamblados y provistos de una interfase web sencilla. Se adoptó PHP[23] 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[24] [W3C, 1999] [Kay, 2001]. La elección de XSLT obedece a criterios técnicos específicos: dado que el recurso lingüístico con 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 presentación en Html.

Se analizaron distintos buscadores encontrando que Google[25] tiene la limitación de 10 palabras por consulta, y una estrategia compleja puede llegar a tener muchas palabras más. Se analizó el buscador Yahoo![26] y se observó que no tiene estas limitaciones. Por eso se utilizó este último buscador en las experiencias.

A continuación se describe el prototipo desarrollado tomando como base el modelo de arquitectura propuesto.

El prototipo y su código se encuentran en el CD-ROM adjunto. También puede ser probado en jsaer.openisp.net/ribs1.1.

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.2.1 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.

[22] www.google.com/apis
[23] www.php.net
[24] www.w3.org/TR/xslt
[25] www.google.com
[26] www.yahoo.com

Ingreso del concepto:

En la pantalla de la Figura A.2.1, 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.

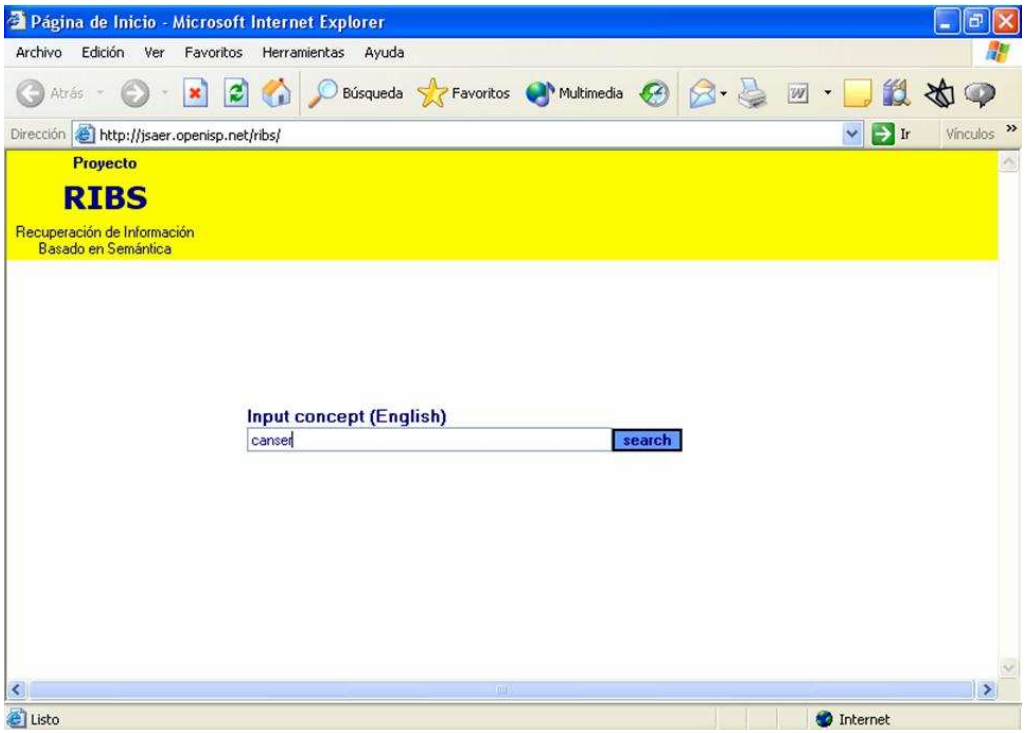


Figura A.2.1. Pantalla inicial del prototipo de refinador semántico

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.2.2.

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.2.2 se muestra la pantalla resultante del corrector ortográfico para este caso.

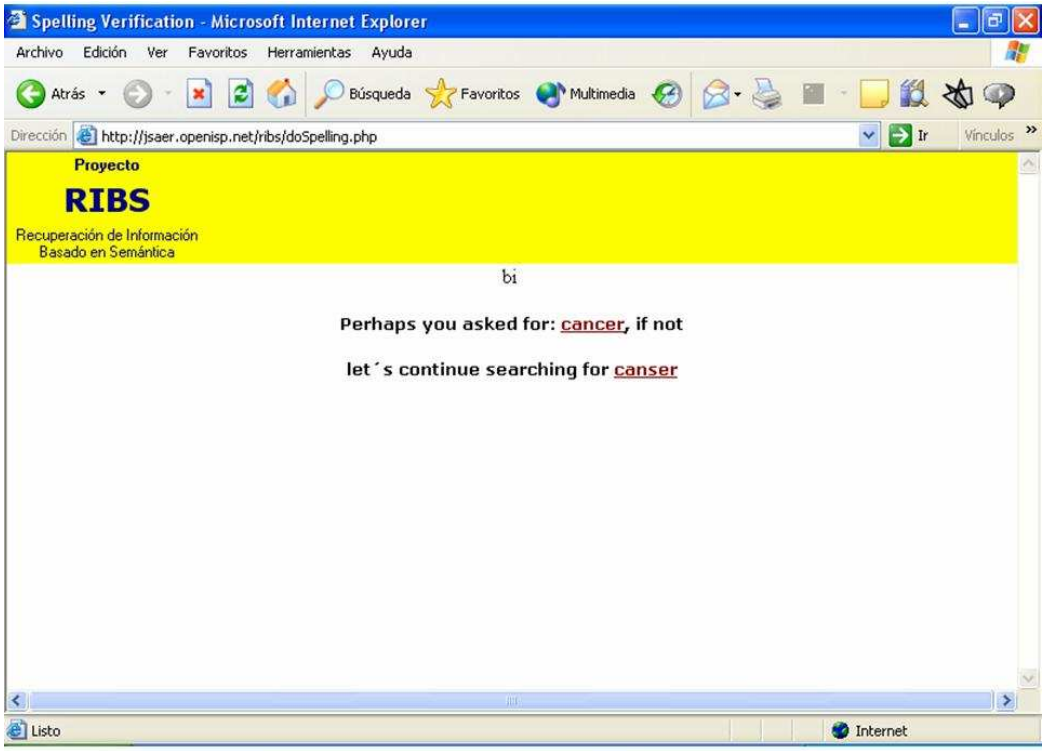


Figura A.2.2. Pantalla resultante de la corrección ortográfica para un término ingresado con errores.

En esta pantalla el usuario debe seleccionar cuál de los dos términos es de su interés: el término sugerido por el corrector ó 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 ésto 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.2.3.

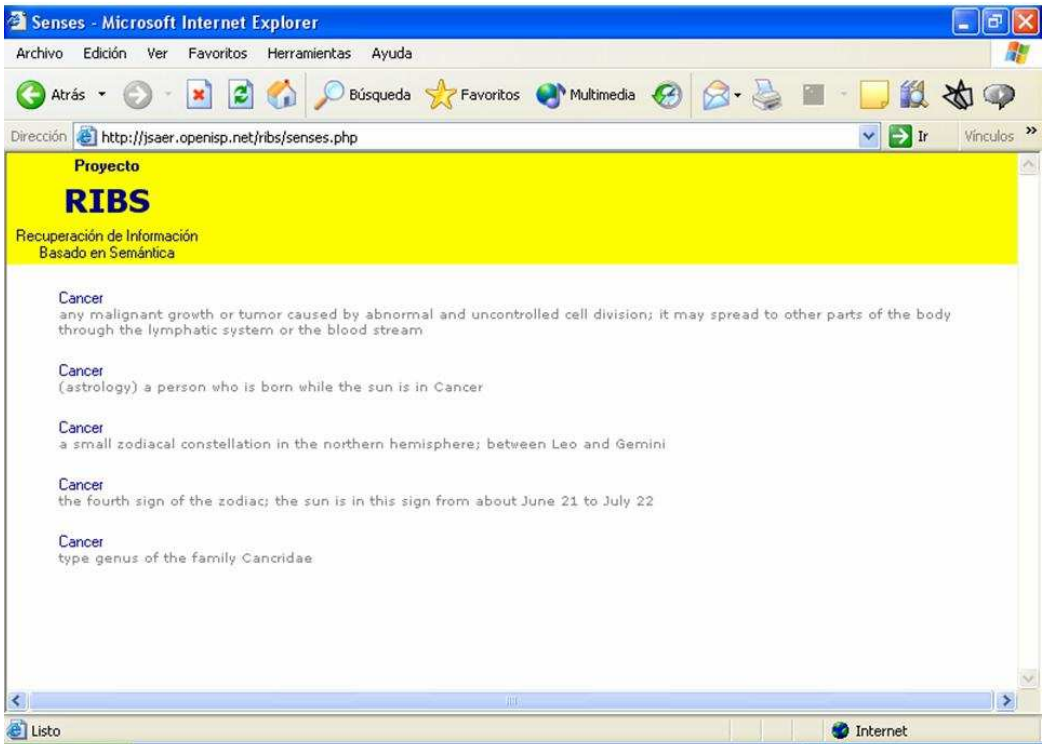


Figura A.2.3. Pantalla que muestra las distintas acepciones de un término.

Continuando con el ejemplo, en la Figura A.2.3, 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.2.4. 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.

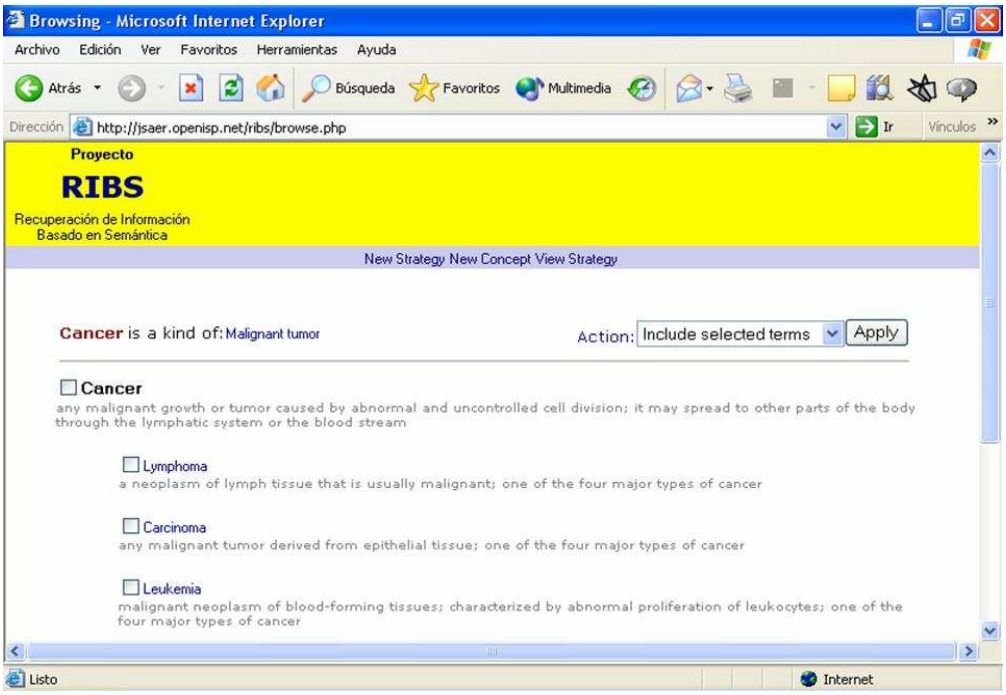


Figura A.2.4. Pantalla que muestra parte de la jerarquía conceptual de "cancer" en medicina.

Continuando con el ejemplo, en la Figura A.2.4, 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.2.5.

En lugar de elegir un hipónimo, el usuario podría elegir el hiperónimo. Es decir, puede ascender ó 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 ó 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.

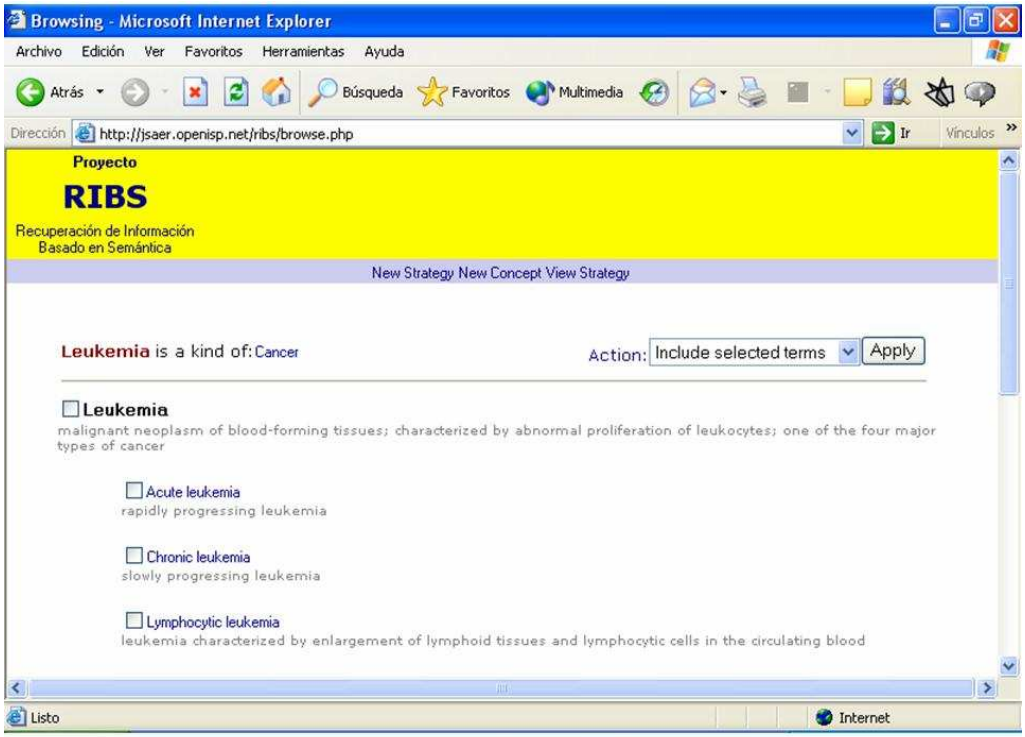


Figura A.2.5. Pantalla que muestra la jerarquía conceptual de "leukemia".

El usuario puede decidir continuar con el término de partida "cancer", ó puede elegir uno, ó 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 ó 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 ó 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.2.5, 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.2.1.

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.2.6. 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.

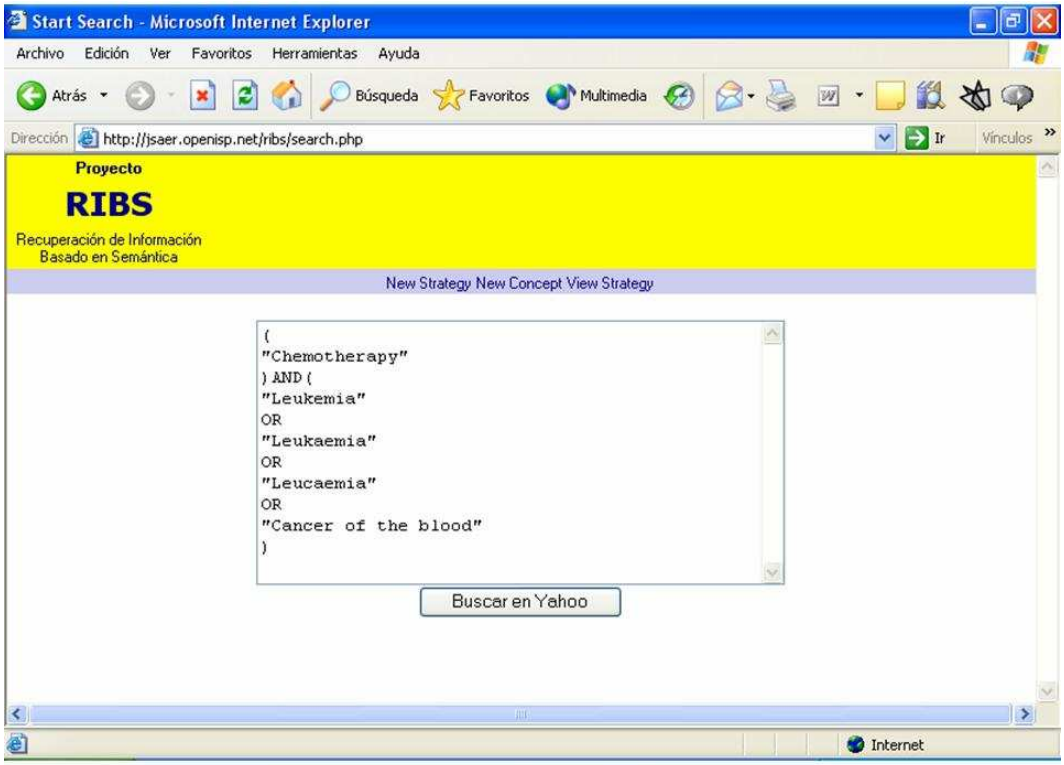


Figura A.2.6: Estrategia resultante de la consulta "quimioterapia utilizada en leucemia"

Para este prototipo, se utiliza actualmente el buscador Yahoo!, pero esta estrategia podría escribirse según las sintaxis de consulta de distintos buscadores ó 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.2.7.

(“Chemotherapy”) AND (“Leukemia” OR “Leukaemia” OR “Leucaemia” OR “Cancer of the blood”)

Figura A.2.7: 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.2.8.

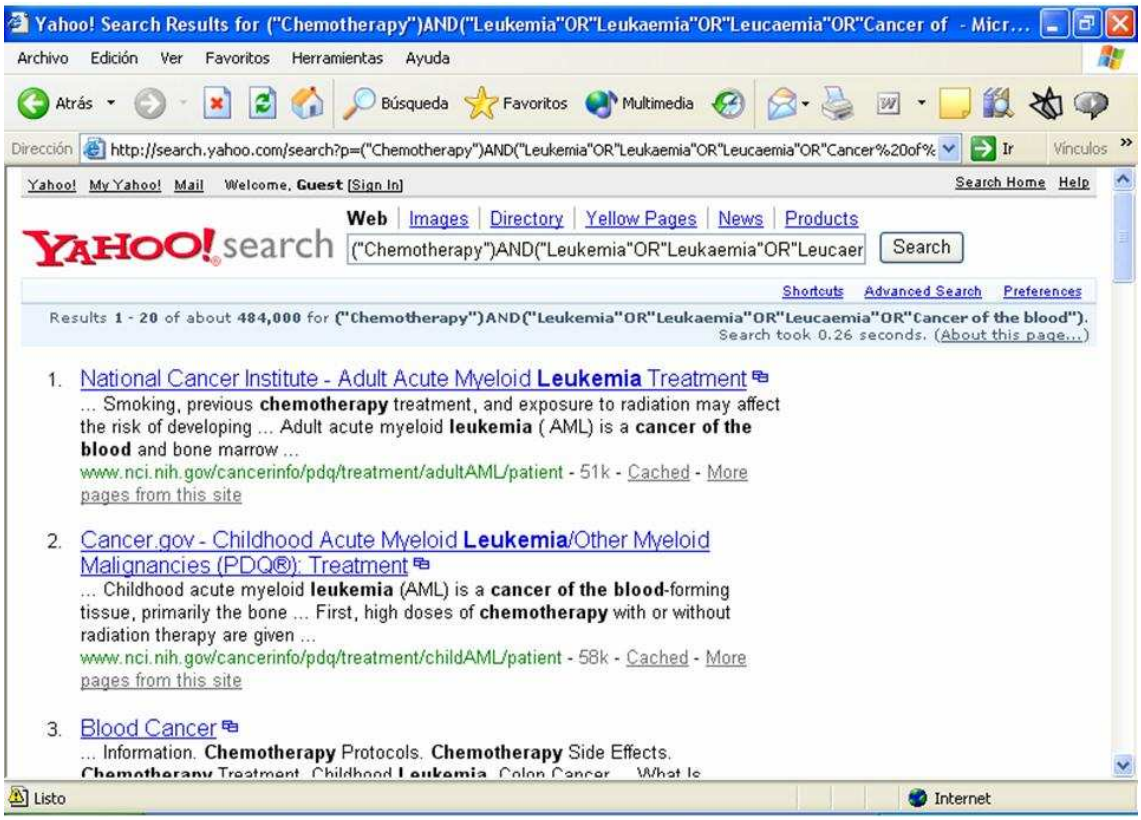


Figura A.2.8: Resultados de la consulta enviada al buscador Yahoo! para el ejemplo “quimioterapia utilizada en leucemia”