Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 21-29 recibido 27-04-2011 aceptado 19-05-2011

¿De verdad sabes lo que quieres buscar? Expansión guiada visualmente de la cadena de búsqueda usando ontologías y grafos de conceptos*

Do you really know what do you want to search? Visually guided expansion of search string using ontologies and concepts graphs

Manuel de la Villa
Depto. de Tecnologías de la Información
Universidad de Huelva
manuel.villa@dti.uhu.es

Sebastián García Pérez
Escuela Politécnica Superior
Universidad de Huelva
sebastian.garcia@alu.uhu.es

Manuel J. Maña
Depto. de Tecnologías de la Información
Universidad de Huelva
manuel.mana@dti.uhu.es

Resumen: Múltiples trabajos hablan de la escasez de términos usados en las cadenas de búsqueda, lo que dificulta que se discriminen eficientemente los documentos de interés del usuario. Los buscadores devuelven miles de documentos recuperados, produciéndose resultados inadecuados, sin una conexión semántica con la consulta y escasamente relacionados con las necesidades del usuario. Este hecho se agrava en un ámbito biomédico donde el paciente no suele dominar el vocabulario especializado necesario para la precisa definición de sus necesidades de información. Presentamos un método de expansión y enriquecimiento de la cadena de búsqueda mediante la creación de un modelo visual esquemático, un grafo de conceptos relacionados semánticamente con la ayuda de ontologías como UMLS y Freebase.
Palabras clave: Recuperación de información, cadena de búsqueda, ontologías, UMLS, Freebase, grafos de contextos.

Abstract: Many reports talk about the shortage of terms commonly used in the search strings, making it difficult to effectively discriminate relevant documents from the user. Search engines return thousands of documents recovered, leading to inadequate results, with no semantic connection with the consultation and little to do with the user's needs. This is heightened in a biomedical field where the patients does not usually dominate the specialized vocabulary needed for the precise definition of their information needs. We present a method of expansion and enrichment of the search string by creating a visual model diagram, a graph of semantically related concepts with the help of ontologies as UMLS and Freebase.
Keywords: Information retrieval, search string, ontologies, UMLS, Freebase, context graphs.

1 Introducción

Con la popularización de Internet en la última década, el volumen de información electrónica accesible ha experimentado un crecimiento exponencial. Esta sobrecarga de información que caracteriza a nuestra sociedad ha otorgado un papel predominante a los sistemas de búsqueda y recuperación de información. En el ámbito biomédico, el problema se agrava si cabe; la documentación disponible en formato electrónico crece y se vuelve más imprescindible, tanto en la formación de los futuros médicos, como en el ejercicio de la profesión y en la divulgación de información a los pacientes. En este dominio, los profesionales y los usuarios en general necesitan herramientas orientadas a proporcionar medios para acceder y visualizar la información de manera adecuada para sus necesidades.

Pero, sin lugar a dudas, uno de los mayores problemas de los sistemas de recuperación de información es la correcta y adecuada

* Este trabajo ha sido parcialmente financiado por ERDF (TIN2009-14057-C03-03)

definición de las necesidades de información del usuario (Jansen, Spink y Koshman, 2007).

Estos problemas se deben, en parte, a que en muchos casos las interfaces de los sistemas de búsqueda no son las adecuadas (Hearst, 1999). Pero también a que cuando el usuario inicia una búsqueda desconoce realmente qué puede serle útil y, por tanto, le resulta complicado especificar las características destacables de los elementos de información potencialmente útiles (Belkin, 2000).

El resultado de este tipo de consultas es habitualmente una lista de miles de documentos recuperados, difícilmente relacionados con las necesidades reales del usuario y de compleja navegación y asimilación.

Nuestro objetivo es ayudar al usuario a definir y concretar la búsqueda mediante la construcción de un grafo que recoja las relaciones semánticas extraídas de una ontología, a partir de un concepto biomédico. El usuario selecciona nuevos términos directamente del grafo, que amplían la cadena de búsqueda, aumentando su capacidad descriptiva, de modo que las necesidades de información expresadas se acerquen lo más posible a las necesidades reales de información.

2 La problemática de la definición de la necesidad de Información

En el problema de la búsqueda de información, las estrategias que utilizan los usuarios para recuperar información podemos dividirlas en dos grandes clases: *querying* (interrogación) y *browsing* (exploración). En el *querying* el usuario introduce en el sistema una serie de palabras clave (representación de sus necesidades de información), tras lo que el sistema le devuelve una lista de resultados pertinentes para su consulta, normalmente ordenada por relevancia.

Los sistemas basados en *querying* resultan útiles y muy eficientes en multitud de casos. Sin embargo, cuando el usuario no tiene completamente claro qué está buscando o, cuando es incapaz o tiene dificultades para formalizar sus necesidades de información a través del lenguaje de consulta, requiere de un modelo alternativo de acceso a la información (Herrero-Solana, Hassan; 2006).

Las consultas que realizan los usuarios de sistemas de Recuperación de Información (RI) tienen habitualmente un carácter tan amplio que difícilmente pueden reflejar una necesidad específica de información. Un informe sobre el comportamiento de los usuarios de motores de búsqueda en la Web (Jansen, Spink y Koshman, 2007) indica que el número medio de palabras por consulta es 2.85, de forma que más del 50% de la consultas contiene un máximo de dos términos y casi el 18,5% de las mismas es de un único término.

La constatación de la dificultad de la elección de las palabras adecuadas para representar sus necesidades de información, de formular consultas que se muestren efectivas recuperando información relevante ha suscitado un gran interés por las técnicas de modificación de consultas (para una revisión de dichas técnicas puede consultarse, por ejemplo, (Baeza-Yates y Ribeiro-Neto, 1999) o (Efthimiadis, 1996)).

2.1 Ayudas a la mejora de de la consulta.

Como se acaba de introducir, podemos resumir que las técnicas que tratan de mejorar la consulta lo hacen de dos formas posibles: cambiando el peso de los términos, es decir, su importancia relativa, y añadiendo nuevos términos. La información para mejorar la consulta puede provenir de distintas fuentes:

- Realimentación por relevancia,
- Análisis local o global, de documentos recuperados y de la colección y
- la utilización de recursos de Procesamiento del Lenguaje Natural (PLN) como diccionarios, tesauros o bases de datos léxicas (Mandala et al., 2000).

La reformulación de la consulta puede ser llevada a cabo automáticamente por el sistema o puede realizarse de forma que el usuario retenga el control de la interacción. En este caso, el sistema sugiere una lista de posibles términos con los que expandir la consulta, de forma que el usuario reformula manualmente la consulta. Experimentos con usuarios muestran las preferencias de estos por mantener el control sobre la forma en que se reformula la consulta (Belkin et al., 2001).

Podríamos establecer varias clasificaciones, atendiendo a los posibles métodos de ayuda a la definición de la necesidad de información:

1) Atendiendo al momento en que se realiza podríamos hablar de ayudas **pre-recuperación** (p.ej. tesauros o diccionarios para expansión de la cadena de búsqueda) **y post-recuperación** (agrupamiento de resultados, selección de términos que etiqueten cada grupo, de la realimentación por relevancia, etc.).
2) Según su presentación en la interfaz, podríamos hablar de **ayudas visuales y no visuales**. La primera, que es la que nos interesa, ha dado lugar a un área de investigación, los estudios enfocados al diseño de Interfaces Visuales para la Recuperación de Información o VIRIs.

2.2 Expansión basada en ontologías

El principal objetivo de la expansión de la consulta es añadir nuevos términos significativos a la consulta inicial. Los nuevos términos que resulten del método de selección deben proporcionar información contextual con el fin de mejorar los resultados de recuperación. Una de las maneras más recientes de obtención de la información contextual es derivada de modelos de conocimiento tales como las ontologías.

El problema con las técnicas tradicionales de expansión es que están dirigidas por el contenido. El contenido del corpus se analiza para extraer los términos candidatos a la ampliación de la consulta. Esto sólo puede funcionar si hay suficientes documentos para trabajar y también que estos documentos contienen un conjunto razonable de los términos que representan el área objeto de la consulta. Los modelos de conocimiento independientes del corpus no sufren de este inconveniente.

2.2.1 Uso de ontologías de dominio general

Las ontologías también se han utilizado para ayudar a la expansión de consultas desde principios de los noventa con un éxito desigual. La ontología de mayor uso y difusión ha sido WordNet (Miller, 1995), una gran base de datos léxica del inglés caracterizada por agrupar sustantivos, verbos, adjetivos y adverbios en conjuntos de sinónimos (*synsets*), cada uno expresando un concepto distinto. Su uso más habitual ha sido como herramienta de desambiguación.

En (Bhogal et al., 2007) se hace una completa revisión al estado del uso de ontologías en la expansión de consultas. En Gonzalo et al. (1998)se observa que si las consultas no son desambiguadas, la indexación por *synsets* es, al menos, tan buena como la indexación estándar de texto, llegando a mejorarla en más de un 29%. De Buénaga et al. (1997) también prueba que la expansión de texto añadiendo términos de *synsets* de las categorías en que se encuentran incluidos, mejora el rendimiento del sistema.

Voorhees (1993) concluye que en consultas con pocos términos puede ser difícil desambiguar porque la jerarquía IS-A no es suficiente para seleccionar de forma fiable el sentido correcto del sustantivo. Por el contrario, la ampliación de consultas aporta poca diferencia en la eficacia de recuperación para las consultas con cadenas largas.

Finkelstein et al. (2002) describen un sistema de búsqueda basado en el contexto que desambigua en las consultas mediante extracción semántica de palabras clave y agrupamiento para generar nuevas consultas. Los resultados muestran que el uso de contexto para guiar un proceso de búsqueda del usuario ofrece claras mejoras.

Navigli y Velardi (2003) argumentan que la expansión con sinónimos e hiperónimos tiene un efecto limitado sobre el rendimiento de recuperación de información web. Ellos sugieren que otro tipos información semántica de una ontología es más eficaz como las glosas y nodos comunes. Usando WordNet y Google crean una red semántica para cada sentido de la palabra y se califican según los nodos comunes. Los resultados del experimento mostraron una mejora sistemática sobre la consulta sin expandir.

En (Song et al., 2007) se propone un algoritmo híbrido de expansión de consultas que combina reglas de asociación con ontologías y técnicas de PLN. Safar and Kefi (2003) presentan un método de expansión de consultas basado en la ontología del dominio y la estructura entramado.

2.2.2 Uso de ontologías de dominio específico

En general, el problema que plantean las ontologías independientes del dominio, como WordNet, radica en que debido a su amplia cobertura, los términos ambiguos tienden a ser

problemáticos. Para tareas de búsqueda muy concretas, las ontologías específicas del dominio son la elección más adecuada.

Estas ontologías se construyen en diferentes ámbitos de aplicación como derecho, medicina, arqueología, agricultura, geografía, multimedia, negocios, economía, historia o incluso noticias de prensa, por nombrar algunos.

Fu, G. et al. (2005) presentan técnicas de ampliación de consultas basadas tanto en un dominio como en una ontología geográfica, dando como resultado la mejora en las búsquedas.

En la TREC Genomics Track, Hersh et al. (2003) realizaron un experimento sobre la base de sinónimos de nombre de gen y otro experimento evaluó la expansión de consultas utilizando recursos de conocimiento externos.

Nilsson et al. (2005) usan una ontología de dominio específico propia (SUiS) para llevar a cabo la ampliación de consultas en un sistema de respuesta a la pregunta. Los tipos de preguntas se limitan a quién, qué, donde y cuando. Los sinónimos e hipónimos sólo se utilizan para aumentar la precisión. Los experimentos han demostrado una mejora en los resultados.

Díaz-Galiano et al. (2009) describen el uso de la ontología médica MeSH para mejorar un sistema de recuperación mediante la expansión de la consulta con términos MeSH.

Huang (2000) usa directorios web como ontologías. El directorio jerárquico permite al usuario buscar en el portal relevante para esa categoría, cualquier ambigüedad se puede resolver antes de recuperar documentos.

Magnini y Speranza (2002) mezclan una ontología global con una ontología lingüística especializada, en este caso una ontología de la economía.

3 *Interfaces para Visualización en recuperación de Información*

La visualización aplicada a la información ayuda a las personas a formar una imagen mental del espacio informativo. Si la visualización tiene lugar en una interfaz cuyo objetivo es la recuperación de información, la expresión acuñada para este tipos de sistemas es VIRIs (*Visual Information Retrieval Interface*). A continuación se analizan una serie de VIRIs basados en metáforas visuales de gran popularidad y con numerosas implementaciones prácticas en la Web.

Los **Tree Maps** son representaciones planas bidimensionales de estructuras de datos jerárquicas, en los que se optimiza el espacio visual ocupándolo completamente (Shneiderman, 1992). Una variante de este tipo son los mapas semánticos autoorganizativos de (Lin et al., 1991), basados en mapas de Kohonen. Las áreas son proporcionales en tamaño a la frecuencia de aparición de los términos y pueden ser agrupadas en diferentes colores. Los términos que co-ocurren con mayor frecuencia se colocan en una misma área o en una cercana. En Figura 1 se muestra un visualizador de dependencias entre conceptos biomédicos.

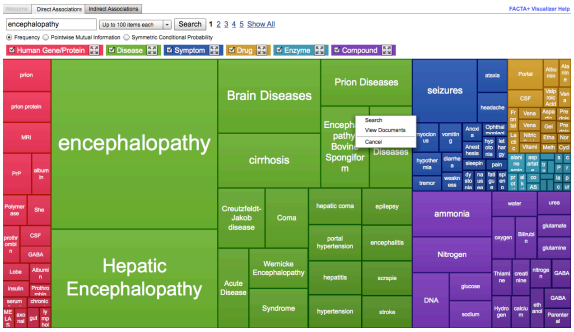

Figura 1. Resultados de búsqueda en FACTA+Visualizer[1].

Los **Tag Cloud** o 'nubes de etiquetas', son un sencillo modelo de VIRI con forma de lista ponderada de palabras clave, que se han popularizado a partir del surgimiento de aplicaciones de software social.

Los **Grafos** permiten expresar de una forma visualmente muy sencilla y efectiva las relaciones que se dan entre elementos de muy diversa índole (Dursteler, 2004), y resultan especialmente útiles en la representación visual de estructuras de datos en red y jerárquicas. Los grafos son la metáfora visual más extendida y en ella nos vamos a centrar, realizando la siguiente especialización:

1) Basados en Tesauros visuales relacionadas con WordNet®

Los Synsets se interrelacionan por medio de relaciones léxicas y semántico-conceptuales. Dentro de este tipo encontramos a *Snappy*

[1]http://refine1-nactem.mc.man.ac.uk/facta-visualizer/

Words (ver Figura 2), *Visual Thesaurus*[2], *Lexipedia*[3] *o Visuwords*[4], extraen información del recurso para construir un grafo derivado donde el nodo inicial es el concepto buscado, del que se da su definición y salen aristas que expresan relaciones semánticas con nuevos nodos o conceptos.

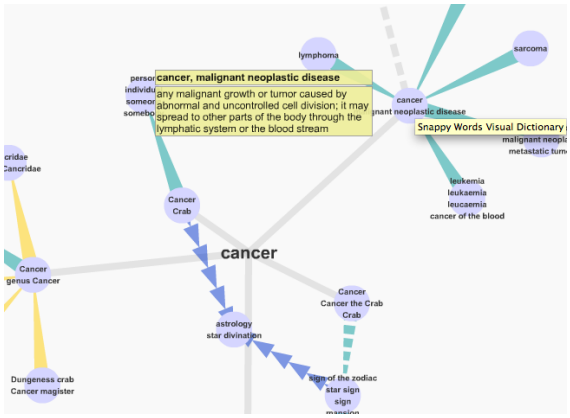

Figura 2. Snappy Words[5].

2) Herramientas de ayuda genérica

En este amplio grupo de herramientas de dominio general incluimos las que no usan WordNet.

Dos recopilaciones de propuestas clásicas de interfaces de visualización pueden encontrarse en (Herrero-Solana y Hassan, 2006) y (Marcos, 2005). Plantean metáforas visuales que van desde *Prise* (una espiral siguiendo el orden de relevancia; o con un gráfico de tres ejes, cada uno correspondiente a un descriptor) a *Bead* (espacio tridimensional en que los artículos se comportan como partículas magnéticas donde la fuerza de atracción tiende a ubicar los artículos similares más cerca), *Tilebars* (segmento dividido proporcionalmente en más celdas, cuanto más extenso sea el documento), pasando por *NIRVE* (en 3D, los clusters en forma de cajas de colores que indican el concepto medio de los documentos que lo conforman, y cuyo tamaño es indicador del número de documentos) o *SPIRE* (Spatial Paradigm for Information Retrieval), actualmente IN-SPIRE[6], destacando sus metáforas *Galaxies* (los documentos aparecen como estrellas y constelaciones) y *ThemeView* (documentos como montañas en un paisaje, valles, ausencia de palabras clave). Esta metáfora geográfica es similar a la presentada en *VxInsight*[7].

Eyeployer[8], es un motor gráfico de conocimiento que proporciona un interfaz fácil de usar, un diagrama de sectores, que divide el espacio de resultados en categorías que engloban conceptos que se relacionan con el buscado. *SearchCube*[9], es un motor de búsqueda visual que presenta los resultados de la búsqueda web previsualizados en celdas de caras de un cubo tridimensional.

WikiMindMap[10] construye un mapa mental a partir de la información que encuentra sobre un concepto en la Wikipedia.

Google Wonder Wheel (Figura 3), la Rueda de Búsquedas de Google, propone nuevas búsquedas basándose en la experiencia de usuario y en las relaciones entre conceptos, usando un grafo en media pantalla, mientras actualiza la otra mitad con los documentos más relevantes para los términos seleccionados (una metáfora parecida usa *Yahoo Correlator*[11]).

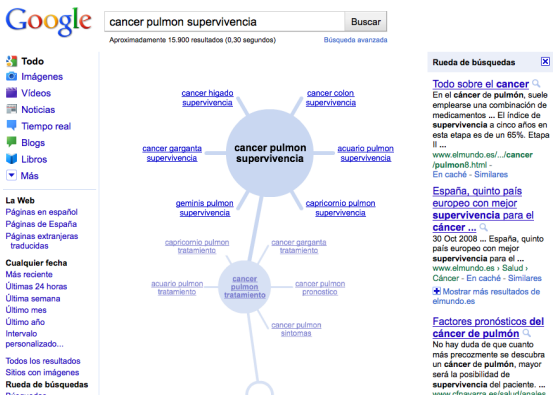

Figura 3. Google Rueda de Búsquedas.

3) Herramientas de ayuda biomédicas.

En este grupo incluiremos aquellos interfaces visuales que, o bien se centran en colecciones de documentos biomédicos, o bien la metáfora

[2] http://www.visualthesaurus.com/
[3] http://www.lexipedia.com/
[4] http://www.visuwords.com/
[5] http://www.snappywords.com/
[6] http://in-spire.pnnl.gov/
[7] http://www.cs.sandia.gov/~dkjohns/JIIS/Vx_Intro.html
[8] http://eyeplorer.com/
[9] http://www.search-cube.com/
[10] http://www.wikimindmap.org/
[11] http://correlator.sandbox.yahoo.net/

visual que sirve de soporte tiene un componente biológico.

NIH Visual Browser[12] aplica el algoritmo DrL (*Distributed Recursive Graph Layout*) para agrupar en clusters documentos de investigación con alto grado de similitud de trabajos financiados por el NIH.

PubAnatomy[13] permite explorar visualmente las relaciones entre las estructuras anatómicas, los procesos fisiopatológicos, los niveles de expresión genética y las interacciones proteína-proteína sobre un esquema del cerebro para una exploración más eficiente de la base de datos de Medline.

4 Descripción del sistema

En los apartados anteriores ha quedado claro que una de las maneras de mejorar la expresividad de la consulta es la adición de nuevos términos, tarea en que era un factor clave la utilización de recursos lingüisticos.

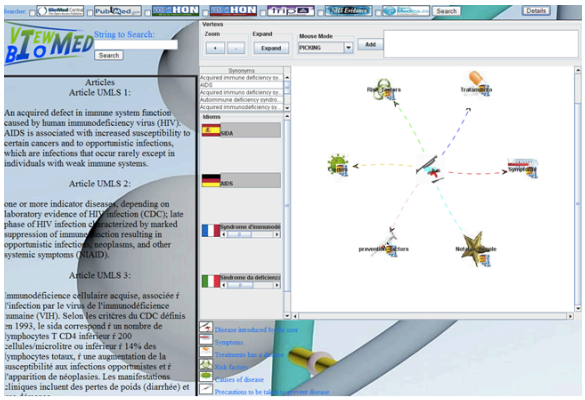

Figura 4. Interface de la aplicación completa con el grafo y definiciones.

Así mismo, también se ha referenciado la preferencia del usuario por mantener el control sobre la reformulación de la consulta. Ese es nuestro desafío, la elaboración de una ayuda visual a la adición de términos en la cadena de búsqueda, en pre-recuperación, controlada por el usuario y guiada por la información contenida en ontologías. Puede observarse la interfaz en la Figura 4.

[12] http://scimaps.org/maps/nih/2007/

[13] http://brainarray.mbni.med.umich.edu/Brainarray/Prototype/Pubanatomy/

4.1 Ontologías usadas en nuestro sistema

En nuestro caso, la utilidad de las ontologías será la de ayudar en la selección del término a buscar, identificación del concepto y extracción de conocimiento semántico.

Vamos a repasar los dos recursos usados en este trabajo, primero el conjunto de recursos que UMLS proporciona y que se combinan para el tratamiento semántico de textos y después FreeBase.

4.1.1 UMLS

Para el procesado semántico, consistente en el análisis e identificación de los conceptos y relaciones subyacentes en un texto, se requiere que el texto pueda ser mapeado a una estructura de conocimiento, como la que en el ámbito biomédico proporciona el proyecto Unified Medical Language System (UMLS). Consiste en tres componentes, el SPECIALIST Lexicon, el Metathesaurus y la UMLS Semantic Network (Rindflesh et. al., 2005).

Usaremos el Metathesaurus para la identificación de los conceptos biomédicos de la cadena de búsqueda. A partir del concepto, sobre un conjunto reducido, multilingüe, de vocabularios, se localiza su representación en otros idiomas (castellano, francés, italiano y alemán), así como sus principales sinónimos (es decir, diferentes términos o variantes léxicas en otros idiomas que son representados con el mismo concepto biomédico)

4.1.2 Freebase

Freebase (Kochhar, Mazzocchi y Paritosh, 2010) es una inmensa base de datos pública que recopila tres clases de información: datos, que incluye información a bajo nivel; texto, que recoge documentación como la descripción de un tópico; y multimedia, que incluye ficheros tales como imágenes, videos y audio. Una entidad o tópico es una persona, lugar o cosa única y singular. Cada tópico tiene su propia página web y es la unidad fundamental.

Al ser su estructura de datos no-jerárquica, Freebase puede modelar relaciones mucho más complejas entre elementos individuales que una base de datos convencional. Estos tipos y las propiedades y los conceptos

relacionados se llaman *Schemas*. En el mundo de la web semántica, ésto se conoce como Ontología.

La información en Freebase se crea, estructura y mantiene colaborativamente. Contiene más de 300 millones de tripletas a través de 12 millones de tópicos que incorporan más de 17000 tipos y 46000 propiedades. Una parte significativa de esta información se obtiene a partir de procesos de minería de datos y aprendizaje que operan sobre fuentes de datos semi-estructurados como la Wikipedia.

Toda esta información es accesible de diversos modos: desde sitios web, aplicaciones y navegadores, ya sea a través de la API, de la propia web de Freebase o aplicaciones de usuario alojadas en la plataforma Acre. MQL (*Metaweb Query Language*) es una API para realizar consultas mediante programación a Freebase. Esto nos permite incorporar conocimiento desde Freebase en nuestras aplicaciones o webs. Es análogo al lenguaje de consulta SPARQL usado en RDF. Usa objetos JSON como consultas a través de peticiones y respuestas HTTP estándar.

El proceso que realiza la aplicación web se puede dividir en tres partes claramente diferenciadas:

4.2 Pre-recuperación del concepto.

En esta fase se realizan principalmente tres tareas:

Tarea 1. *Ayudar al usuario a definir el concepto inicial.* A medida que el usuario va introduciendo la cadena inicial de búsqueda, haciendo uso de Freebase y de AJAX, se ayuda al usuario mostrando una lista desplegable de conceptos (y su tipo) en cuya descripción aparece la cadena introducida. El usuario puede recorrer la lista, mostrándo una pequeña ventana con una entrada resumida del concepto seleccionado (ver Figura 5).

Tarea 2. *Consulta a las fuentes de información*. Una vez que el usuario ha encontrado el concepto del que desea obtener información, se realizan las consultas con ese término tanto a Freebase como a UMLS.

- *Consulta a Freebase:* Utilizamos la API que nos proporciona Freebase, en concreto, una consulta MQL de la información que necesitamos, en concreto obtenemos: *Tratamientos, Causas, Factores de Riesgos, Factores de Prevención, Síntomas y Personas Famosas.*
- *Consulta a UTS* (UMLS Terminology Services), que incluye una API de Servicio Web para, de manera remota vía Internet, consultar y recuperar información de los recursos léxicos y semánticos de UMLS. A través *de SOAP 1.2* y *WDSL* se describe la interfaz pública a dichos servicios Web. UTS nos proporciona la siguiente información: *Sinónimos y Traducción del término a los principales idiomas.*

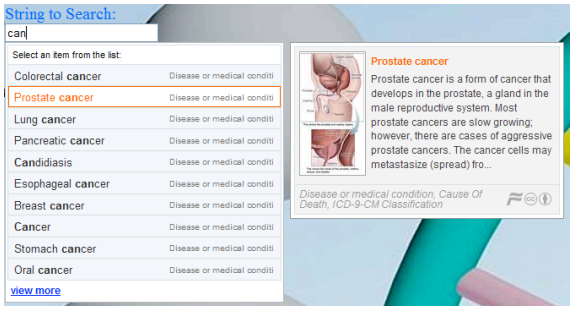

Figura 5. Ayuda en el concepto inicial.

Tarea 3. *Guardar la información en un archivo XML.* Esto se hace como almacenamiento intermedio con vistas a posibles comunicaciones futuras incrustado en otros sistemas.

4.3 Construcción del grafo

Cuando hemos conseguido recuperar toda la información necesaria, se construye el grafo (Figura 6) usando la librería JUNG (O'Madadhain et al., 2005).

Se muestra en el centro del grafo el concepto inicial con aristas a nodos que representan los conjuntos de Tratamientos, Causas, Factores de Riesgo, Síntomas y Celebridades que han padecido la enfermedad, cada uno de ellos con un icono característico.

Inicialmente, cada conjunto se muestra comprimido, para no sobrecargar la interfaz y el usuario decide que conjunto expandir.

4.4 Navegación por el grafo

Llegados a este punto, la interfaz permite al usuario de manera intuitiva: hacer zoom, hacer *panning*, cambiar disposición de nodos, expandir los nodos grupales (mostrando en

detalle, p.ej., un nodo nuevo por cada factor de riesgo asociado a la enfermedad inicial), poder ver la información asociada a cada nodo en las diferentes ontologías que utilizamos, poder añadir términos desde un nodo a la cadena de búsqueda, etc.

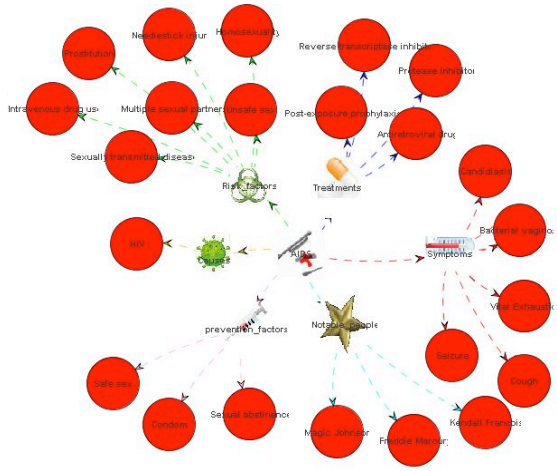

Figura 6. Imagen del grafo expandido.

4.5 Recuperación con la cadena expandida

El sistema le da la posibilidad al usuario de hacer una búsqueda más detallada en buscadores médicos más especializados. Se muestra en la parte superior de la interfaz una selección de buscadores para que el usuario escoja en cuales quiere hacer dicha búsqueda. Cada búsqueda generará una ventana del navegador por cada buscador seleccionado con el resultado de la consulta de la cadena expandida.

5 Conclusiones

En este trabajo se ha introducido el problema de la habitualmente escasa definición de las necesidades de información del usuario y el uso de ontologías en la expansión de consultas como medio para conseguir una recuperación de información eficiente.

Se ha realizado un estudio del arte de las metáforas visuales más difundidas en el campo de los interfaces de visualización para recuperación de información. Podríamos concluir que si el usuario no logra crear un modelo mental adecuado del sistema, la representación que se le ofrece no habrá cumplido su objetivo. Se hace necesario continuar desarrollando aplicaciones que mejoren esta comunicación del sistema con el usuario.

Se ha presentado una propuesta de interfaz visual de pre-recuperación, disponible online[14], basada en un grafo de conceptos construido con información semántica obtenida de dos ontologías, Freebase y el Metathesauro UMLS.

Quedaría pendiente un trabajo de evaluación del sistema y de la interfaz orientado a usuarios, donde usuarios realizaran búsquedas con esta interfaz frente a otras tradicionales, para medir el grado de mejora de la eficiencia (una mejor expresión de la necesidad proporcionará mayor índice de documentos relevantes) así como satisfacción del usuario.

Por último, indicar que esta prevista su integración en un sistema de recuperación de información propio (de la Villa et al., 2010) para su uso tanto en pre-recuperación como en post-recuperación para clasificar y navegar por los documentos devueltos.

Bibliografía

Baeza-Yates, R. y Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. Capítulo 5. Addison-Wesley Longman Publishing Company, New York.

Belkin, N. J. 2000. Helping People Find What They Don't Know. *Communications of the ACM*, 43(8):58-61.

Belkin, N., Cool, C., Kelly, D., Lin, S. J., Park, S. Y., Perez-Carballo, J. y Sikora, C. 2001. Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Information Processing & Management*, 37(3):403-434.

Bhogal, J., Macfarlane, A., Smith, P. (2007). A review of ontology based query expansion. *Information Processing & Management*, 43(4), 866-886. Pergamon Press, Inc.

De Buenaga M.,, Gómez-Hidalgo J.M., Díaz-Agudo,B. 1997. Using WordNet to complement training in formation in text categorization. Proceedings of RANLP-97. Tzigov Chark, Bulgary.

de la Villa, M., Muñoz A., Millán M., Maña, M. 2010. A Biomedical Information Retrieval System based on Clustering for

[14] http://www.uhu.es/manuel.villa/viewmed/

Mobile Devices. *BioSEPLN10, Workshop on Language Technology applied to biomedical and health documents*. SEPLN 2010 (Valencia).

Efthimiadis, E. 1996. Query expansion. *Annual Review of Information Science and Technology (ARIST)*, 31:121-187

Finkelstein L., Gabrilovich E., Matias Y., Rivlin E., Solan Z., Wolfman G, y Ruppin E. 2002. Placing search in context: the concept revisited. *TOIS*, 20(1), 116–131.

Fu, G. et al. 2005. Ontology-Based Spatial Query Expansion in Information Retrieval ODBASE: OTM Confederated International Conferences.

Gonzalo, J. et al. 1998. Indexing with WordNet synsets can improve text retrieval *Coling-ACL 98*.

Hearst, M. A. 1999. User Interfaces and Visualization. En R. Baeza-Yates y B. Ribeiro-Neto, eds., *Modern Information Retrieval* , págs. 257-323. Addison-Wesley Longman Publishing Company, New York.

Herrero-Solana, V., Hassan, Y. 2006. Metodologías para el desarrollo de Interfaces Visuales de Recuperación de Información: análisis y comparación. *Information Research*, 11(3)

Hersh, W., Bhupatiraju, R. T., y Price, S. 2003. Phrases, Boosting, and Query Expansion Using External Knowledge Resources for Genomic Information Retrieval. *TREC* (503–509).

Huang, L. 2000. A survey on web information retrieval technologies. In *ECSL*. New York.

Jansen, B. J., Spink, A. and Koshman, S. 2007. Web searcher interaction with the Dogpile.com metasearch engine. *Journal of the American Society for Information Science and Technology*, 58: 744–755.

Kochhar, S., Mazzocchi, S. y Paritosh, P. 2010. The Anatomy of a Large-Scale Human Computation Engine. *KDD-HCOMP '10*, Washington, DC (USA).

Lin, X., Soergel, D. y Marchionini, G. 1991. A Self-organizing Semantic Map for Information Retrieval. Proc. *ACM Int. SIGIR '91*.

Magnini, B., y Speranza, M. 2002. Merging global and specialized linguistic ontologies. In *Proceedings of the workshop Ontolex-2002 ontologies and lexical knowledge bases*, LREC-2002 (pp. 43–48).

Mandala, R., Tokunaga, T. y Tanaka, H. 2000. Query expansion using heterogeneous thesauri. *Inf. Process. Manage.* 36(3): 361-378

Marcos, M. C. 2005. Visual elements in search systems and information retrieval. Yearbook Hipertext.net, n. 3 (May 2005).

Miller, G.A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* Vol. 38, No. 11: 39-41.

Navigli, R., y Velardi, P. 2003. An analysis of ontology-based query expansion strategies workshop on adaptive text extraction and mining (ATEM 2003). *In 14th European conference on machine learning (ECML 2003.*

O'Madadhain, J., Fisher, D., Smyth, P., White, S., and Boey, Y.-B. 2005. Analysis and visualization of network data using JUNG. *Journal of Statistical Software ,* VV:1-35.

Rindflesh, T.C., Fiszman, M., Libbus, B. 2005. Semantic interpretation for the biomedical research literature. Capítulo 14. *Medical Informatics. Knowledge Management and Data Mining in Biomedicine*. Springer's Integrated Series in Information Systems.

Safar, B., Kefi, H. 2003. Domain ontology and Galois lattice structure for query refinement. *Proceedings of the 15th IEEE international conference on tools with artificial intelligence*, Sacramento, California, pp. 597–601.

Shneiderman, B. 1992. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, v. 11 n. 1, p.92-99, Jan. 1992.

Song M., Song I-Y., Hu X., Allen R.B. 2007. Integration of association rules and ontologies for semantic query expansion. *Data & Knowledge Engineering* 63. 63–75

Voorhees, E. 1993. Using wordnet to disambiguate word senses for text retrieval. *ACM SIGIR*, 171–180.