Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 39-46 recibido 03-05-2011 aceptado 22-05-2011

Estudio del uso de Ontologías para la Expansión de Consultas en Recuperación de Imágenes en el Dominio Biomédico*

Using Ontologies for Query Expansion in Image Retrieval in the Biomedical Domain

Jacinto Mata, Mariano Crespo, Manuel J. Maña
Dpto. de Tecnologías de la Información. Universidad de Huelva
Ctra. Huelva - Palos de la Frontera s/n. 21819 La Rábida (Huelva)
{jacinto.mata, mariano.crespo, manuel.mana}@dti.uhu.es

Resumen: La existencia de grandes colecciones de imágenes médicas ha generado un interés cada vez mayor por el acceso a este tipo de información. En este artículo abordamos este problema desde el punto de vista de la recuperación basada en la información textual relacionada con la imagen. La hipótesis inicial planteada es que la expansión de las consultas podría mejorar la efectividad de un sistema de recuperación de imágenes. Se han utilizado diferentes elementos de información contenidos en las ontologías MeSH y UMLS. La expansión se ha llevado a cabo tanto a nivel de término como de concepto. Para la experimentación se ha utilizado la colección de documentos ImageCLEF del año 2009. Los resultados obtenidos muestran un ligero incremento de la medida MAP y una diferencia más significativa cuando la evaluación se realiza usando la medida F. La conclusión final es que la expansión de consulta no es suficiente para conseguir una mejora sustancial de la efectividad en este tipo de sistemas de recuperación de información.
Palabras clave: Recuperación de imágenes basada en texto, dominio médico, expansión de consultas, ontologías.

Abstract: The existence of huge collections of medical images in scientific repositories and hospital databases has generated increasing interest in the access to this information. In this paper we address this problem focusing on image retrieval based on textual information related to the image. The initial hypothesis is that query expansion could improve the effectiveness of image retrieval systems. In this proposal, we have used several information elements contained in MeSH and UMLS ontologies. The expansion has been carried out at both term and concept levels. For the experiment we have used the document collection ImageCLEF 2009. The results show a slight increase in MAP and a more significant difference when the evaluation is performed using the F-measure. The final conclusion is that the query expansion is not sufficient to achieve a substantial improvement in the effectiveness of this type of information retrieval systems.
Keywords: Text-based image retrieval, medical domain, query expansion, ontologies.

1 Introducción

En la actualidad existe una gran cantidad de información biomédica en formato electrónico. Son ya numerosas las grandes colecciones de datos médicos con información visual y textual disponible para investigadores, profesionales de la salud y, en general, para todas las personas interesadas en este tipo de información.

En general, los sistemas de recuperación de información han centrado sus esfuerzos en mejorar la accesibilidad a la información textual. Sin embargo, existe un creciente interés en optimizar el acceso a la información visual. La recuperación de imágenes (*image retrieval*) es un campo menos explorado que la recuperación de texto. Básicamente hay dos enfoques para abordar el problema de la

* Este trabajo ha sido parcialmente financiado por el Ministerio de Ciencia e Innovación, el Plan E del Gobierno Español y la Unión Europea con cargo al FEDER (TIN2009-14057-C03-03)

recuperación de imágenes. El enfoque basado en el contenido visual de la imagen (*content-based image retrieval)* y el basado en la información textual relacionada con la imagen (*text-based image retrieval*). En este artículo presentamos un trabajo basado en el enfoque textual.

El principal problema en la recuperación de imágenes a partir de su información textual asociada radica en la dificultad de realizar una correcta anotación del contenido de la imagen. En muchas ocasiones resulta difícil expresar con palabras lo que ocurre en una imagen (situaciones, sentimientos, etc.). Además, a esto hay que añadir las dificultades propias del lenguaje (sinonimia, hiperonimia, hiponimia, uso de abreviaturas, etc.).

En la mayoría de las ocasiones, el uso de recursos externos mejora ostensiblemente el rendimiento de los sistemas de recuperación de información. Actualmente se hace un gran esfuerzo para desarrollar estos recursos, especialmente en el ámbito biomédico, con el fin de ayudar al usuario final a manejar estos grandes volúmenes de información (Bodenreider, 2004) (Nelson, Johnston y Humphreys, 2004). Entre los recursos más utilizados se encuentran las ontologías. En el ámbito de la biomedicina existe un amplio número de tesauros y ontologías. Entre ellas se encuentran GO[1] (Stevens, Goble y Bechhofer, 2000), MeSH[2] (Nelson et al., 2001) y UMLS[3] (Bodenreider, 2004).

En este artículo se presenta un estudio sobre el uso de dos de las ontologías más utilizadas en el ámbito médico (MeSH y UMLS) para la expansión de consultas con el objetivo de mejorar un sistema de recuperación de imágenes médicas. La colección empleada para el estudio y la experimentación que mostramos en este artículo es la que se utilizó en la tarea de recuperación de imágenes médicas de la competición ImageCLEF[4] del año 2009 (Müller et al., 2010). En este corpus, la información asociada a cada imagen (metadatos) está formada por la leyenda de la imagen y el título del artículo al que pertenece. También incluye el identificador del artículo para poder acceder al texto completo.

El resto del artículo se estructura de la siguiente forma. En la sección 2 se describen los trabajos relacionados más relevantes. En la sección 3 se presenta la colección de documentos utilizada para el estudio y experimentación. Posteriormente, en la sección 4, se describen las ontologías empleadas. En la sección 5 se describen las estrategias de expansión desarrolladas con cada una de las ontologías empleadas. En la sección 6 se muestran y discuten los resultados obtenidos en los distintos experimentos y, finalmente, las conclusiones y trabajos futuros se detallan en la sección 7.

2 Trabajos relacionados

Actualmente existen diversos sistemas de recuperación de imágenes y sus documentos asociados. La mayoría de ellos están especialmente diseñados para trabajar en el dominio biomédico. Algunos ejemplos de estos sistemas son *Yale Image Finder* (Xu, McCusker y Krauthammer, 2008), *ARRS Goldminer* (Kahn y Thao, 2007) y *BioText* (Hearst et al., 2007).

Estos sistemas obtienen las imágenes de los documentos a partir del texto contenido en los títulos, en las leyendas de las figuras, en los resúmenes o en el documento completo. Sin embargo, debido a las propias características de las colecciones, en numerosas ocasiones estos sistemas pierden eficacia y, por consiguiente, no ofrecen los resultados esperados.

Existen diversos trabajos donde se presentan estudios sobre el efecto de la utilización de las ontologías MeSH y UMLS para la expansión de consultas. En (Lu, Kim y Wilbur, 2009), los autores investigan una estrategia de expansión de consultas haciendo uso del proceso avanzado de búsqueda que proporciona PubMed[5] denominado *Automatic Term Mapping* (ATM). Para realizar el estudio utilizan una colección de 64 consultas y unas 160.000 citas de MEDLINE que fueron las utilizadas en las TREC Genomics Track[6] de los años 2006 y 2007. Entre los resultados destacan un aumento en la medida F del 21.5% y 23.3% en las colecciones del 2006 y 2007 respectivamente cuando hacen expansión de consultas. Los autores concluyen que la expansión de consulta mediante MeSH en PubMed puede mejorar la efectividad de la recuperación pero que, en

[1] http://www.geneontology.org/
[2] http://www.nlm.nih.gov/mesh/meshhome.html
[3] http://www.nlm.nih.gov/research/umls/
[4] http://www.imageclef.org/2009
[5] http://www.ncbi.nlm.nih.gov/pubmed/
[6] http://ir.ohsu.edu/genomics/

situaciones reales, la mejoría puede no resultar significativa para los usuarios de PubMed.

En el trabajo de (Díaz et al., 2009a), los autores detallan los experimentos realizados en la tarea de recuperación de información médica del ImageCLEF del año 2008. Crearon tres tipos de colecciones diferentes, la primera con las leyendas de las imágenes y los títulos de los artículos, la segunda con las leyendas, los títulos y el texto de la sección donde se referencia a la imagen, y el tercero con el texto completo del artículo. Realizaron expansión de consultas con las ontologías MeSH y UMLS, obteniendo los mejores resultados con la expansión de consultas usando MeSH e indexando las leyendas y títulos de los artículos. La mejora del MAP respecto a su línea base fue del 12.5%.

Los mismos autores, en (Díaz, Martín y Ureña, 2009b), presentan los resultados obtenidos realizando expansión de consultas mediante los términos de entrada (*entry terms*) de la ontología MeSH utilizando las colecciones de la tarea de recuperación de información médica del ImageCLEF de los años 2005 y 2006. En este trabajo combinan la información textual con la información visual y concluyen que se produce una mejora en la eficacia de la recuperación usando, de forma conjunta, ambas estrategias. Concretamente, obtienen una mejora en la medida MAP del 25.07% y del 37.3% respecto a su *baseline* en las colecciones de los años 2006 y 2005 respectivamente.

La principal diferencia entre los trabajos previos y la propuesta que presentamos en este artículo consiste, básicamente, en la selección de los elementos de información utilizados para realizar la expansión de las consultas. Nuestro objetivo era explorar algunas de las relaciones que aportan estas ontologías para establecer un conjunto de estrategias de expansión.

3 Descripción de la colección

La colección de imágenes seleccionada para llevar a cabo la evaluación del sistema ha sido proporcionada por la organización del ImageCLEF. En concreto, en este trabajo se ha utilizado la de la edición de 2009. Ese año, la colección estaba compuesta por 74.902 imágenes y la participación fue muy amplia, ya que se registraron 38 grupos de investigación de todo el mundo.

ImageCLEF proporciona a los participantes de la competición dos ficheros en formato XML con la colección de documentos y con las consultas (*topics*). Para la evaluación de los sistemas existe un tercer fichero de texto con los juicios de relevancia de las imágenes para cada consulta (*ground truth*).

El fichero XML de la colección de imágenes está formado por 74.902 registros (un registro por cada imagen). Los metadatos de cada registro lo forman el identificador de la imagen, su URL, la leyenda de la imagen, el título del artículo al que pertenece, el identificador del artículo en PubMed, la URL del artículo y el nombre del fichero de la imagen. En la Figura 1 se muestra un ejemplo de registro de este fichero.

```
<record>
<figureID>27979</figureID>
<figureURL>http://radiology.rsnajnls.org/cgi/content/full/210/1/11/F1</figureURL>
<caption> Figure 1.Illustration of a neonate at autopsy whose demise was attributed to "thymic death." The caption drew attention to the "enormous size of the thymus," which is actually normal in appearance.(Reprinted, with permission, from reference 6.)
</caption>
<title>The right place at the wrong time: historical perspective of the relation of the thymus gland and pediatric radiology
</title>
<pmid>9885579</pmid>
<articleURL>http://radiology.rsnajnls.org/cgi/content/full/210/1/11</articleURL>
<imageLocalName>27979.jpg</imageLocalName>
</record>
```

Figura 1: Ejemplo de registro de la colección de imágenes

```
1   0   227775   0
1   0   227776   0
1   0   198482   0
....
2   0   129252   1
2   0   53729    0
2   0   50647    1
....
25  0   66087    0
25  0   187404   0
25  0   125313   0
```

Figura 2: Fichero de juicios

El fichero de juicios está formado por cuatro columnas. La primera columna indica el número de la consulta. La tercera representa el identificador de la imagen. La cuarta columna indica la relevancia (un valor binario donde 0 representa *no relevante* y 1 indica *relevante*).

La segunda columna se ignora para este tipo de evaluación. En la Figura 2 se muestra el formato del fichero de juicios.

El fichero de las consultas está formado por 25 registros (un registro por consulta). La información proporcionada para cada uno de los registros está formada por el identificador de la consulta, el tipo y el texto de la consulta formulada en inglés junto con su traducción en francés y alemán. Para este trabajo únicamente se han utilizado las consultas en inglés. En la Figura 3 se muestra un registro de este fichero.

```
<topic>
        <ID>2</ID>
        <TYPE>visual</TYPE>
        <EN_DESCRIPTION>
                Breast cancer mammogram
        </EN_DESCRIPTION>
        <FR_DESCRIPTION>
                Mammographies d'un cancer du sein
        </FR_DESCRIPTION>
        <DE_DESCRIPTION>
                Mammogramm mit Brustkrebs
        </DE_DESCRIPTION>
</topic>
```

Figura 3: Ejemplo de registro del fichero de consultas

4 Ontologías MeSH y UMLS

Las ontologías representan el conocimiento de un dominio concreto en forma de un conjunto de conceptos y de relaciones entre ellos. En el dominio biomédico son muchos los recursos terminológicos y ontológicos existentes y también son variadas sus aplicaciones en PLN: recuperación de información, búsqueda de respuestas, resumen automático o clasificación, entre otras. Las dos ontologías utilizadas en la experimentación, MeSH y UMLS, son una iniciativa de la *National Library of Medicine* de los EEUU.

MeSH es un vocabulario controlado utilizado para la indexación de artículos en Medline. Está formado por conjuntos de términos, denominados descriptores, organizados en una estructura jerárquica que permite la búsqueda a diferentes niveles de especificidad. En la actualidad, MeSH está compuesto por 26.142 descriptores o *Main Headings*. Este vocabulario es el que se utiliza para indexar las citas de Medline. Las formas alternativas, sinónimos y términos relacionados con estos descriptores se denominan términos de entrada (*entry term*). En MeSH existen más de 177.000 términos de entrada.

UMLS es un conjunto de fuentes de conocimiento y herramientas software para el PLN de textos del dominio biomédico. Las fuentes de conocimiento que forman UMLS son: el Léxico Especializado, el Metatesauro y la Red Semántica. El *Léxico Especializado* describe las características sintácticas de términos en inglés de carácter biomédico y general, proporcionando la base para el PLN en el dominio biomédico. Además de etiquetas con la categoría gramatical, para cada entrada se incluyen variaciones ortográficas que ocurran, inflexiones en nombres, verbos y adjetivos. El *Metatesauro* es una recopilación de más de 100 vocabularios y terminologías médicas, asociando cada término a más de un millón de conceptos semánticos que a su vez se engloban en al menos uno de los tipos de la red semántica. La *Red Semántica* constituye una ontología del más alto nivel de la Medicina, compuesta por 135 tipos semánticos asignados a conceptos del Metatesauro y por 54 tipos de relaciones entre los mencionados tipos semánticos.

5 Expansión de consultas

El término *expansión de consultas* se utiliza cuando, en un motor de búsqueda, se añaden nuevos términos a la consulta del usuario con el objetivo de aumentar la eficacia en la recuperación. En los últimos años, los sistemas basados en expansión de consultas que están mejorando ostensiblemente sus resultados son los que hacen uso de recursos externos como ontologías o jerarquías léxicas.

En el ámbito de la recuperación de la información (RI), las ontologías son utilizadas con frecuencia. Entre sus principales aplicaciones se encuentran la expansión de consultas, la indexación semántica de los documentos y la organización de los resultados de la búsqueda. Sin embargo, debido a la gran cantidad de información que ofrecen, en la mayoría de las ocasiones no resulta sencillo su utilización en las tareas de RI. En (Jimeno, Berlanga y Rebholz, 2010), los autores proponen un algoritmo para refinar ontologías para tareas de recuperación de información.

En las siguientes secciones se describen las estrategias de expansión utilizadas para el estudio y experimentación.

5.1 Estrategias de expansión haciendo uso de MeSH

La ontología MeSH ofrece numerosas posibilidades para realizar la expansión de los términos de la consulta. Por ejemplo, en (Díaz, Martín y Ureña, 2009b), los autores utilizan los términos de entrada (*entry term*) como criterio para llevar a cabo la expansión de las consultas.

En el trabajo que presentamos se han utilizado otros elementos que ofrece la ontología. Concretamente se ha hecho uso de dos tipos de referencias cruzadas (*SeeRelatedDescriptor* y *ConsiderAlso*) y de la propia estructura jerárquica en la que MeSH organiza sus descriptores.

La expansión de la consulta se realiza a nivel de término. Los términos que son descriptores en MeSH se expanden con el contenido del elemento *SeeRelatedDescritor* o *ConsiderAlso*. El primero de ellos asocia el descriptor con otros descriptores relacionados mediante una referencia cruzada. El objetivo de este tipo de asociaciones es proporcionar otros descriptores que pueden ser más apropiados para un caso particular. Por ejemplo, el término *Cancer* se expande, entre otros, con los términos *Carcinogens*, *Antibodies*, *Neoplasm*, *Genes* y *Tumor Suppressor*.

El elemento *ConsiderAlso* referencia a otros descriptores relacionados mediante raíces lingüísticas. Por ejemplo, el término *Brain* referencia a las raíces lingüísticas *CEREBR-* y *ENCEPHAL-*.

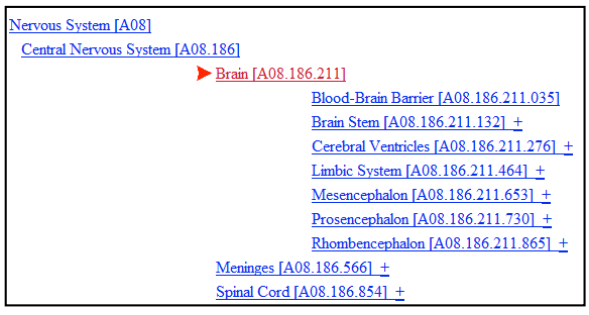

Figura 4: Extracto del árbol MeSH

La tercera estrategia de expansión está basada en la propia estructura de árbol en la que MeSH organiza sus descriptores. En este caso, si el descriptor es un nodo padre se expande con sus descriptores hijos. Si el descriptor no tiene hijos no se produce expansión. En la Figura 4 se muestra un breve extracto del árbol MeSH donde se observa que el descriptor *Brain* tiene siete hijos mientras que el descriptor *Central Nervous System* tiene tres.

En muchas ocasiones un descriptor está formado por más de un término, por lo que realizar la expansión a nivel de término no es lo más eficaz. Por ejemplo, si en la consulta *Pituitary Adenoma* se tratara cada término de forma independiente, la palabra *Pituitary* no corresponde con ningún descriptor. Sin embargo, la unión de los dos términos sí corresponde a un descriptor ya que *"Pituitary Adenoma"* es un concepto biomédico.

La última estrategia de expansión consiste en identificar los conceptos médicos formados por más de un término que aparecen en la consulta. Si este concepto es un descriptor se expande con los descriptores hijos. Además, para cada término de la consulta se comprueba si es un descriptor y, en caso afirmativo, se expande con el mismo criterio.

5.2 Estrategias de expansión haciendo uso de UMLS

La ontología UMLS es aún mucho más amplia que MeSH. De hecho, MeSH es uno de los vocabularios que forma parte del Metatesauro de UMLS. En este trabajo se han utilizado dos tipos de relaciones que UMLS realiza entre sus conceptos:

- RN: Relación estrecha entre conceptos.
- RB: Relación amplia entre conceptos.

En este caso hacemos uso de los conceptos en lugar de los términos y se han desarrollado dos estrategias de expansión. En la primera se expanden los conceptos mediante RN y en la segunda, la expansión se realiza mediante los conceptos relacionados por RB. Por ejemplo, si la consulta es *breast cancer mammogram,* ésta se transforma en: *breast cancer "breast cancer" mammogram*, teniendo en cuenta que *"breast cancer"* es un concepto biomédico. La consulta sólo se expande si *"breast cancer"* tiene alguna relación RN o RB con algún otro concepto. Siguiendo con el ejemplo, *"breast cancer"* se relaciona, mediante el tipo RN, con el concepto *"Mammary Neoplasms"* y, por tanto, la consulta final será: *breast cancer "breast cancer" "Mammary Neoplasms" mammogram.*

La diferencia entre la expansión usando esta ontología y la utilizada en MeSH es que, en ésta, sólo se expanden los conceptos mientras

que con MeSH se expanden tanto los conceptos cómo los términos.

6 *Resultados y discusión*

En esta sección se detallan los experimentos que se han llevado a cabo para evaluar las distintas estrategias de expansión.

Como es habitual en las competiciones de ImageCLEF, la medida más relevante para la organización fue el *Mean Average Precision* (MAP). Se trata de una medida de un solo valor que tiene en cuenta la posición en el *ranking* de los documentos relevantes recuperados. Se calcula como la media de las precisiones obtenidas cada vez que se recupera un documento relevante. Para una colección de necesidades de información, se promedia como se indica en (1).

$$MAP = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{r_j}\sum_{k=1}^{r_j}\text{Precisión}(R_k) \qquad (1)$$

Donde:

m: Número de consultas.
r_j: Número de documentos relevantes en la consulta j.
R_k: *k-ésimo* documento relevante recuperado, en el *ranking* de documentos recuperados.

Para realizar la evaluación del sistema, se utilizó el software TREC_EVAL, desarrollado por el TREC (*Test REtrieval Conference*), el cual permite calcular las medidas más utilizadas para evaluar la eficacia en la recuperación.

Para evaluar la influencia de la expansión de las consultas en el resultado de la recuperación, se diseñó el siguiente *baseline*. Se construyó un índice con las leyendas de todas las imágenes de la colección y se lanzaron las 25 consultas tal como las proporcionó la competición, realizando un preprocesado consistente en la eliminación de palabras vacías y en la reducción a la raíz. Para realizar este proceso se utilizó el motor de búsqueda *Lucene*[7] sin realimentación por pseudo-relevancia.

En las siguientes secciones se muestran los resultados obtenidos con las estrategias de expansión descritas anteriormente para cada una de las ontologías estudiadas. Las mejoras obtenidas con las técnicas de expansión propuestas están por debajo de lo esperado. Sin embargo, estos resultados no son comparables con los trabajos previos (Díaz et al., 2009a), (Díaz, Martín y Ureña, 2009b) ya que las colecciones de evaluación son distintas.

6.1 Experimentación con MeSH

En la Tabla 1 se muestra la medida MAP obtenida por cada consulta con cada una de las estrategias de expansión, donde:

- C: Número de consulta.
- BL: *BaseLine*.
- SRD: *SeeRelatedDescriptor*.
- CA: *ConsiderAlso*.
- AT: Árbol MeSH con términos.
- AC: Árbol MeSH con conceptos.

C	BL	SRD	CA	AT	AC
1	0,1064	0,1064	0,1064	0,0963	**0,1404**
2	0,3278	**0,2020**	0,2001	0,1903	0,1632
3	0,3311	**0,3311**	**0,3311**	**0,3311**	**0,3311**
4	0,2238	0,2649	0,2238	0,2902	**0,3208**
5	0,4644	0,4644	0,3492	0,4644	**0,5295**
6	0,3231	0,3168	0,3231	0,0229	**0,3379**
7	0,3062	**0,3062**	0,3014	**0,3062**	**0,3062**
8	0,5794	0,5794	0,5794	0,5794	**0,5986**
9	0,3347	0,2415	0,1937	0,1558	**0,3347**
10	0,3010	0,2694	**0,3010**	0,2231	**0,3010**
11	0,4646	0,2515	**0,4646**	0,1311	0,3985
12	0,4777	0,4777	0,4777	**0,5269**	**0,5269**
13	0,0185	0,0018	0,0009	0,0017	**0,0043**
14	0,7629	0,7629	0,7629	0,1298	**0,8620**
15	0,5257	**0,5257**	**0,5257**	**0,5257**	**0,5257**
16	0,1652	0,1651	**0,1652**	0,1557	**0,1652**
17	0,0168	0,0109	0,0085	0,0112	**0,0168**
18	0,7178	0,6602	**0,7178**	0,2894	**0,7178**
19	0,6108	**0,6108**	**0,6108**	0,6020	**0,6108**
20	0,1320	**0,1320**	**0,1320**	0,0447	0,0887
21	0,1217	**0,1217**	**0,1217**	**0,1217**	**0,1217**
22	0,3591	0,1823	0,3443	0,1525	**0,3814**
23	0,1038	0,1038	0,1038	0,2335	**0,2877**
24	0,3720	0,3720	0,3720	0,006	**0,4001**
25	0,3331	**0,3331**	**0,3331**	**0,3331**	**0,3331**
VM	0,3391	0,3117	0,3220	0,2620	**0,3522**

Tabla 1: Resultados de la medida MAP para cada consulta

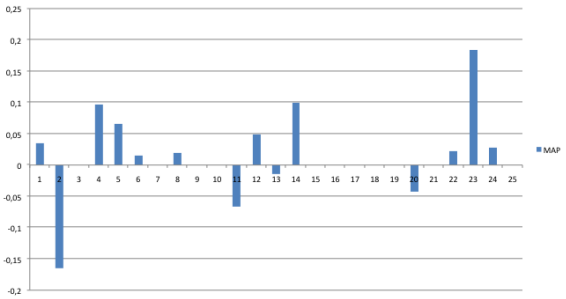

Figura 5: Histograma para las 25 consultas usando la estrategia de expansión AC

[7] http://lucene.apache.org/

La técnica de expansión con la que se obtuvieron mejores resultados fue la de expandir con conceptos médicos a partir del árbol MeSH, obteniendo un valor medio para el MAP de 0,3522, lo que supone un incremento de un 4% respecto a nuestro *baseline*.

En la última fila se muestran los valores medios (VM) para el *baseline* y las 4 estrategias utilizadas. Cabe destacar que en la edición del 2009, el MAP del sistema ganador fue de 0,43 y que una medida de 0,35 se hubiera situado entre las 12 mejores ejecuciones.

En la Figura 5 se muestra el histograma para cada consulta. Se puede apreciar que 10 de las 25 consultas mejoran en esta medida mientras que en 4 de ellas el MAP disminuye, y en 11 permanece igual debido a que no se ha producido expansión. Este hecho nos hace pensar que resultaría conveniente realizar una expansión alternativa cuando no se produzca expansión mediante esta estrategia. Cabe resaltar el decremento tan significativo que se produce en la consulta 2, que hace disminuir considerablemente el valor medio del MAP.

Aunque en el ImageCLEF la medida utilizada para evaluar los resultados fue el MAP, en este trabajo se realizó un estudio completo de las técnicas de expansión desarrolladas, y por ello además del MAP, se calcularon la precisión, la cobertura y la medida F. En la Tabla 2 se muestran los valores obtenidos con cada una de las estrategias de expansión.

Medida	BL	SRD	CA	AT	AC
Precisión	0,1051	0,1035	**0,1048**	0,0593	0,0899
Cobertura	0,6801	0,6611	0,6700	0,6307	**0,6875**
F	0,1185	0,1160	0,1182	0,1017	**0,1319**

Tabla 2: Valor medio de precisión, cobertura y medida F con cada estrategia de expansión

Como se puede observar, la técnica que obtiene el mejor resultado en medida F es también la del árbol MeSH con conceptos médicos. En este caso, el incremento es del 11.3% respecto a nuestro *baseline*.

6.2 Experimentación con UMLS

En la Tabla 3 se muestra el MAP obtenido por cada consulta con las dos estrategias de expansión haciendo uso de UMLS.

En este caso, ninguna de las dos estrategias logró superar la medida MAP de nuestro *baseline*, aunque se obtuvieron valores muy similares. La estrategia de expansión utilizando el tipo de relación RB obtuvo los mejores resultados, con un valor medio para el MAP de 0,3239.

De la misma forma que con las estrategias basadas en MeSH, también se realizó un estudio de las medidas de precisión, cobertura y medida F. En la Tabla 4 se muestran los resultados. Se puede observar que, haciendo uso de las relaciones de tipo RB, se consigue incrementar la medida F consiguiendo una mejora del 5,7%.

Q	BL	RN	RB
1	0,1064	0,0806	**0,1404**
2	**0,3278**	0,1488	0,2669
3	**0,3311**	**0,3311**	**0,3311**
4	**0,2238**	**0,2238**	**0,2238**
5	0,4644	**0,4885**	0,3303
6	0,3231	0,2454	**0,3384**
7	**0,3062**	0,3049	**0,3062**
8	0,5794	0,5684	**0,6174**
9	**0,3347**	**0,3347**	**0,3347**
10	0,3010	**0,3798**	0,3711
11	**0,4646**	**0,4646**	0,4081
12	0,4777	0,2919	**0,4814**
13	**0,0185**	0,0064	0,0051
14	**0,7629**	**0,7629**	**0,7629**
15	**0,5257**	**0,5257**	**0,5257**
16	0,1652	**0,1770**	0,1652
17	**0,0168**	0,0133	0,0147
18	**0,7178**	0,6369	0,7045
19	**0,6108**	0,5704	0,5797
20	**0,1320**	0,0861	0,0783
21	0,1217	0,1217	**0,6108**
22	0,3591	**0,3696**	0,3347
23	0,1038	0,0170	**0,1102**
24	0,3720	**0,3753**	0,2107
25	**0,3331**	0,3036	**0,3331**
Valor medio	**0,3391**	0,3132	0,3239

Tabla 3: Resultados de la medida MAP en cada consulta.

Medida	BL	RN	RB
Precisión	0,1051	0,0637	0,1040
Cobertura	0,6801	0,6786	**0,6829**
F	0,1185	0,1096	**0,1253**

Tabla 4: Valor medio de precisión, cobertura y medida F con cada estrategia de expansión.

7 Conclusiones y trabajo futuro

En este artículo hemos presentado un estudio sobre la utilización de diversas estrategias de expansión de consultas haciendo uso de dos de las ontologías más utilizadas en el ámbito médico, con el objetivo de mejorar la eficacia de un sistema de recuperación de imágenes basado en el contenido textual.

Hemos elegido distintos elementos de las ontologías para realizar la expansión. Los resultados de nuestros experimentos demuestran que no todas las estrategias de expansión consiguen mejorar la efectividad del sistema. Sin embargo, se ha demostrado que haciendo uso de la estructura jerárquica con la que MeSH dispone sus descriptores y expandiendo por términos y conceptos médicos, se puede conseguir una ligera mejora tanto en el MAP como en la medida F.

Con este estudio hemos podido comprobar la dificultad que entraña encontrar una estrategia adecuada para realizar la expansión de consultas. Estamos convencidos que existen elementos o combinaciones de elementos con los que realizar la expansión de las consultas y que puedan mejorar sustancialmente un sistema de recuperación.

En trabajos futuros seguiremos investigando con los diferentes elementos de información de las ontologías MeSH y UMLS. También utilizaremos más información textual asociada a cada imagen. En este sentido, pretendemos añadir al índice las secciones donde aparecen descritas las imágenes en el artículo junto con las leyendas. En este trabajo sólo se ha experimentado expandiendo las consultas. En futuros estudios realizaremos también la expansión de los conceptos médicos que aparecen en el texto con el que se construye el índice. Finalmente exploraremos las posibilidades de la indexación conceptual usando UMLS.

Referencias bibliográficas

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research,* 32(2004) 267–270.

Díaz, M.C., M.A. García, M.T. Martín, L.A. Ureña y A. Montejo. 2009a. Query Expansion on Medical Image Retrieval: MeSH vs. UMLS. *Evaluating Systems for Multilingual and Multimodal Information Access*. Lecture Notes in Computer Science. Volumen 5706/2009, 732-735.

Díaz, M.C., M.T. Martín y L.A. Ureña. 2009b. Query expansion with a medical ontology to improve a multimodal information retrieval. *Computers in Biology and Medicine*, 4, 396-403.

Hearst, M., A. Divoli, H. Guturu, A. Ksikes, P. Nakov, M.A. Wooldridge y J. Ye. 2007. BioText Search Engine: beyond abstract search. *Bioinformatics* 23(16): 2196-2197.

Jimeno, A., R. Berlanga y D. Rebholz. 2010. Ontology refinement for improved information retrieval. *Information Processing & Management*, 46(4), 426-435.

Kahn, C.H. Jr. y C. Thao. 2007. GoldMiner: A Radiology Image Search Engine. *American Journal of Roentgenology* 188:1475-1478.

Lu, Z., W. Kim y W. Wilbur. 2009. Evaluation of query expansion using MeSH in PubMed. *Information Retrieval*, Vol. 12, No. 1, pp. 69-80.

Müller, H., J. Kalpathy–Cramer, I. Eggel, S. Bedrick, S. Radhouani, B. Bakke, C.E. Kahn y W. Hersh. 2010. Overview of the CLEF 2009 Medical Image Retrieval Track. *Lecture Notes in Computer Science*, Volume 6242/2010, 72-84.

Nelson, S.J., D. Johnston y B.L. Humphreys. 2001. Relationships in medical subject headings. *Relationships in the Organization of Knowledge*. Kluwer Academic Publishers, pp.171–184.

Nelson, S.J., M. Schopen, A.G. Savage, J.L. Schulman y N. Arluk. 2004. The MeSH translation maintenance system: structure, interface, design and implementation. M. Fieschi, et al. (Ed.). *Proceedings of the 11th World Congress on Medical Informatics*, pp.67–69.

Stevens, R., C.A. Goble y S. Bechhofer. 2000. Ontology-based knowledge representation for bioinformatics. *Brief Bioinformatics* 1(4), pp. 398-414.

Xu, S., J. McCusker y M. Krauthammer. 2008. Yale Image Finder (YIF): a new search engine for retrieving biomedical images. *Bioinformatics* 24(17): 1968-1970.