

# Análisis de la expansión de consulta para colecciones médicas utilizando información mutua, ganancia de información y la ontología MeSH \*

## *Analysis of the query expansion for medical collections using mutual information, information gain and MeSH ontology*

José M. Perea-Ortega  
Manuel C. Díaz-Galiano

Arturo Montejo-Ráez  
Miguel Á. García-Cumbreras

Sistemas Inteligentes de Acceso a la Información (SINAI)  
Escuela Politécnica Superior, Universidad de Jaén, E-23071 - Jaén  
{jmperea, amontejo, mcdiaz, magc}@ujaen.es

**Resumen:** En este trabajo se muestran diferentes experimentos relacionados con la expansión de consultas en sistemas de recuperación de información médica, utilizando ImageCLEFmed como marco de evaluación. Se han evaluado diferentes técnicas de expansión de consulta como PRF y otras basadas en la utilización de la ontología médica MeSH. En concreto, para este último caso, se han probado diferentes tipos de filtrado de los términos MeSH a expandir, basándonos en una selección previa de categorías MeSH y en la aplicación de conceptos como ganancia de información e información mutua. Los resultados obtenidos demuestran que la expansión PRF mejora levemente el caso base sin expandir la consulta, mientras que la conveniencia de la expansión con términos MeSH es difícil de determinar. No obstante, se puede deducir que la expansión con términos MeSH introdujo demasiado ruido durante el proceso de recuperación, al menos mediante las técnicas aplicadas en este trabajo.

**Palabras clave:** Expansión de consulta, información mutua, ganancia de información, MeSH, ImageCLEFmed

**Abstract:** In this paper we show several experiments related to query expansion in medical information retrieval systems, using ImageCLEFmed as a evaluation framework. We have evaluated different query expansion techniques such as PRF and others based on the use of the medical ontology MeSH. Specifically, in this latter case, different types of filtering have been tested, based on a previous selection of MeSH categories and the application of concepts such as information gain and mutual information. The results show that the PRF expansion slightly improves the base case without expanding the query, while the desirability of the expansion with MeSH terms is difficult to determine. However, we can deduce that the expansion with MeSH terms introduced too much noise during the retrieval process, at least by the techniques applied in this work.

**Keywords:** Query expansion, mutual information, information gain, MeSH, ImageCLEFmed

## 1. Introducción

En los últimos años es ingente la cantidad de información en el dominio médico que podemos encontrar en formato digital. Esta gran

cantidad de información demanda sistemas de procesamiento automático y sistemas de búsqueda efectiva que permitan mejorar el trabajo de sus usuarios (Karamanis, 2007; Müller et al., 2006). Son diversas las técnicas que se pueden aplicar para intentar aproximar el lenguaje utilizado en las colecciones (informes médicos, artículos médicos, etc.) y en las consultas del usuario, entre las cuales destacan la expansión de términos y el filtra-

\* Este trabajo ha sido cofinanciado por el Fondo Europeo de Desarrollo Regional (FEDER), proyecto TIN2009-13391-C04-02 (MICINN), proyecto GeOasis (P08-TIC-41999) de la Junta de Andalucía, proyecto UJA2009/12/14 (Universidad de Jaén) y por el proyecto Geocaching Urbano (RFC/IEG2010)

do de las colecciones.

En este trabajo presentamos varios experimentos relacionados con la expansión de las consultas. Para llevar a cabo dicha expansión es frecuente utilizar recursos externos de conocimiento como las ontologías médicas. En concreto, para este trabajo se ha utilizado la ontología médica MeSH<sup>1</sup>. Normalmente, en los sistemas de Recuperación de Información (*Information Retrieval*, IR) la expansión a ciegas de la consulta suele introducir demasiado ruido durante el proceso de recuperación. Por ello, es conveniente aplicar algún tipo de filtrado previo a los términos que serán añadidos a la consulta. En este sentido, se han aplicado los conceptos de Ganancia de Información (*Information Gain*, IG) (Sebastiani, 2002) e Información Mutua (*Mutual Information*, MI) para seleccionar qué términos son los que aportan más información y de cuáles se puede prescindir.

Este artículo se ha organizado de la siguiente manera. En la Sección 2 presentamos trabajos relacionados con esta temática. La Sección 3 muestra el marco de trabajo utilizado, colecciones, consultas, sistema de recuperación de información y demás herramientas. En las siguientes secciones mostramos los experimentos llevados a cabo y los resultados obtenidos, analizando dichos resultados. Por último, mostramos las conclusiones y futuras líneas de trabajo.

## 2. Trabajo relacionado

La expansión de consulta para la mejora de sistemas de recuperación es una técnica muy extendida. En el ámbito biomédico existen multitud de ontologías y bases de conocimiento que nos ayudan a realizar esta tarea. Por lo tanto, existen numerosos trabajos que utilizan la expansión de consulta para mejorar un sistema IR médico. Un ejemplo de esta mejora es la que nos muestran Aronson y Rindfleisch (1997), que utilizan el programa MetaMap<sup>2</sup> para asociar conceptos del metatesauro UMLS<sup>3</sup> (*Unified Medical Language System*) con la consulta original. Ellos llegan a la conclusión de que la estrategia óptima es combinar la expansión de consulta con la realimentación. Por otro lado, Hersh, Price, y Donohoe (2000) observan que el metatesauro

UMLS puede beneficiar a un sistema de recuperación de información si se realiza la expansión de un grupo reducido de consultas y, por lo tanto, se debe estudiar cuándo es beneficiosa la expansión de una consulta. Nilsson, Hjelm, y Oxhammar (2005) usan sinónimos e hipónimos de la ontología SUIIS (*Stockholm University Information System*) para realizar la expansión de consultas. Los experimentos muestran un incremento de la precisión. Lana-Serrano, Villena-Román, y González-Cristóbal (2008) realizan varios experimentos expandiendo consultas y colección con la ontología MeSH y UMLS, etiquetando las entidades médicas tanto en la colección documental como en las consultas. Con este paso pretenden realizar un proceso similar a la extracción de raíces, pero a nivel de conceptos médicos. Además, crean un diccionario de términos médicos a partir de un subconjunto de UMLS. Posteriormente, utilizan la ontología MeSH para expandir tanto documentos como consultas con los hipónimos de las entidades encontradas. En sus experimentos, realizan expansión de la consulta utilizando MeSH, PRF (*Pseudo Relevance Feedback*) con varias configuraciones e incluso ambos métodos de forma combinada, sin que ningún tipo de expansión consiga superar a dicho caso base. Gobeill et al. (2009) también utilizan la ontología MeSH para expandir tanto la colección como las consultas. Para expandir la colección selecciona los cinco mejores descriptores de MeSH que devuelve un clasificador. Dependiendo del tipo de experimento el clasificador puede tomar como entrada el pie de imagen o el artículo completo donde aparece la imagen. Para expandir la consulta, utiliza el mismo clasificador y algunas reglas manuales. También realizan otro tipo de expansión de la consulta incluyendo los sinónimos del descriptor de MeSH. Con estos experimentos llegan a la conclusión de que la mejor forma de expandir la colección es utilizando el artículo completo como entrada del clasificador. Por otro lado, concluyen que expandir la consulta con los sinónimos de MeSH es demasiado agresivo, ya que se obtienen peores resultados. Por último, Díaz-Galiano (2011) realiza una expansión básica utilizando términos de la ontología MeSH para mejorar un sistema de recuperación de casos médicos. En su estudio concluye que, aunque la expansión mejora de forma global el sistema, los términos añadidos en algunas

<sup>1</sup><http://www.nlm.nih.gov/mesh>

<sup>2</sup><http://metamap.nlm.nih.gov>

<sup>3</sup><http://www.nlm.nih.gov/research/umls>

de las consultas no consiguen aportar más valor semántico y, por lo tanto, en dichos casos, no se mejoran los resultados obtenidos con la consulta sin expandir.

### 3. Marco de trabajo

CLEF<sup>4</sup> (*Cross Language Evaluation Forum*) es un foro de evaluación que aboga por el uso y desarrollo de aplicaciones para la gestión y manejo de librerías digitales. Para ello, desarrollan infraestructuras de prueba, mejora y evaluación de sistemas de recuperación de información multimodal y multilingüe. Para poder mejorar los sistemas se realizan competiciones anuales de evaluación, con el objetivo de crear una comunidad de investigadores y desarrolladores que están interesados en la misma tarea. De esta forma, se facilitan futuras colaboraciones entre grupos con intereses similares. Dentro de las competiciones CLEF podemos encontrar la tarea ImageCLEF<sup>5</sup>. Esta tarea se encarga de evaluar distintos aspectos de los sistemas de recuperación de información multimodales, y más concretamente de aquellos que mezclan información visual y textual. Está compuesta por varias subtarefas clasificadas según el tipo de imágenes que se utilizan, existiendo algunas que trabajan con imágenes genéricas, médicas o extraídas de la Wikipedia<sup>6</sup>.

La tarea concreta sobre recuperación de imágenes médicas dentro de ImageCLEF comenzó su andadura en 2005, y se conoce comúnmente como ImageCLEFmed. En 2008, después de tres años utilizando y ampliando las subcolecciones de consulta, creando nuevas consultas cada año y evaluando los resultados de los sistemas competidores, los organizadores de esta subtarea decidieron crear una gran colección de test (Hersh, Müller, y Kalpathy-Cramer, 2009). El objetivo de dicha colección fue consolidar los recursos ofrecidos en las competiciones del 2005 al 2007, creando una colección de referencia para todos los investigadores que desearan probar y evaluar sus algoritmos de recuperación de imágenes médicas. Además, los esfuerzos realizados se encaminaron a crear una colección de documentos, imágenes y consultas que representaran lo más fielmente posible a las colecciones utilizadas en el mundo real.

<sup>4</sup><http://www.clef-campaign.org>

<sup>5</sup><http://www.imageclef.org>

<sup>6</sup><http://www.wikipedia.org>

Junto con la colección de documentos e imágenes se distribuyeron un total de 85 consultas seleccionadas de las campañas de evaluación entre 2005 y 2007, es decir, 25 consultas del 2005, 30 del 2006 y otras 30 del 2007. Las consultas se proporcionaron en tres idiomas: inglés, francés y alemán. Además, dichas consultas están clasificadas como visuales, textuales o mixtas dependiendo de en qué tipo de sistemas de recuperación se comportan mejor. En la Figura 1 se puede observar una consulta clasificada como visual y compuesta por un texto en tres idiomas junto con tres imágenes.

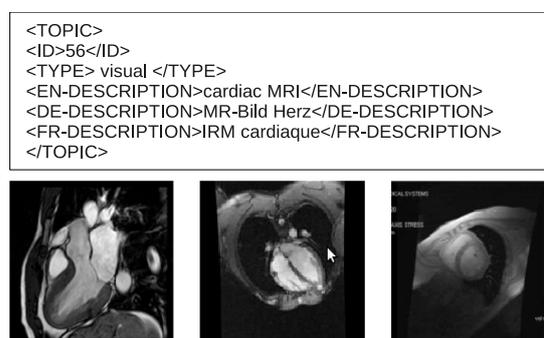


Figura 1: Ejemplo de una consulta de la colección consolidada de ImageCLEFmed

Otro recurso incluido en la colección consolidada son los juicios de relevancia para cada consulta. Los juicios están expresados en tres niveles de relevancia: relevante, parcialmente relevante y no relevante. El número de juicios de relevancia para cada consulta está comprendido entre 800 y 1.200. En definitiva, para los experimentos llevados a cabo durante este trabajo se ha utilizado el marco de evaluación de ImageCLEFmed, compuesto por 85 consultas diferentes.

Por otro lado, para llevar a cabo el proceso de recuperación de información hemos utilizado la plataforma Terrier<sup>7</sup>. Terrier es un motor de búsqueda flexible, eficiente y eficaz, de código abierto, escrito en Java y desarrollado por la Universidad de Glasgow. Esta herramienta implementa los últimos esquemas de pesado y funcionalidades en recuperación de información, además de permitir la indexación automática de colecciones de test conocidas como las proporcionadas por las conferencias TREC y CLEF. El esquema de pesado utilizado en Terrier ha sido el modelo que trae por defecto llamado *inL2*, un modelo ba-

<sup>7</sup><http://terrier.org>

sado en la frecuencia inversa de documento (IDF) que aplica dos tipos de normalización: la primera basada en la sucesión de Laplace y la segunda basada en la frecuencia del término. Por otro lado, Terrier también permite realizar expansión de consulta de forma automática proporcionando diferentes modelos de expansión. El mecanismo de expansión de consulta de Terrier extrae automáticamente los  $m$  términos con más información de los  $n$  primeros documentos recuperados y los utiliza como términos de expansión de la consulta. Durante este proceso, los términos de los  $n$  primeros documentos recuperados son pesados utilizando un esquema basado en el paradigma *Divergence From Randomness* (DFR). La versión 2.2.1 utilizada de Terrier ofrecía tres modelos de expansión de consulta: *Bose-Einstein 1* (Bo1), *Bose-Einstein 2* (Bo2) y *Kullback-Leibler* (KL). Durante la experimentación de este trabajo se ha utilizado el modelo Bo1 por ser el que mejores resultados obtuvo. Con respecto al procesamiento llevado a cabo con los documentos de la colección y las consultas, se han eliminado las palabras vacías y se ha extraído el lema del resto de términos aplicando el algoritmo de Porter (1980).

Por último, las medidas utilizadas para evaluar los experimentos han sido las típicas utilizadas en RI. La precisión (*Precision*,  $P$ ) se define como la proporción de documentos relevantes recuperados sobre el número total de documentos recuperados, la cobertura (*Recall*,  $R$ ) es la proporción de documentos relevantes recuperados del total de documentos relevantes que hay en la colección para esa consulta, y la precisión media (*Mean Average Precision*, MAP) que se considera un buen indicador del comportamiento general del sistema:

$$MAP = \frac{\sum_{d=1}^N P(d)}{N} \quad (1)$$

donde  $N$  es el número de documentos relevantes recuperados,  $d$  es un documento relevante recuperado y  $P(d)$  es la precisión del documento  $d$  recuperado.

#### 4. Experimentos y resultados

El objetivo de este trabajo es analizar cómo se comporta la expansión de consulta en un marco de trabajo como ImageCLEFmed. Para ello, se han llevado a cabo diversos experi-

$n$ docs	$m$ terms	$P$	$R$	MAP
3	10	0,066	0,525	0,244
5	10	0,066	0,523	0,243
10	10	0,065	0,518	0,239
3	15	0,067	0,528	<b>0,247</b>
5	15	0,066	0,526	0,243
10	15	0,065	0,512	0,237
3	20	0,067	0,528	<b>0,247</b>
5	20	0,067	0,527	0,240
10	20	0,065	0,517	0,236

Tabla 1: Resultados de la expansión PRF utilizando Terrier con el esquema Bo1

mentos en los que se han probado diferentes técnicas. Cada uno de estos experimentos y los resultados obtenidos se exponen a continuación.

##### 4.1. Expansión PRF

El primer tipo de expansión consistió en aplicar un mecanismo de realimentación por pseudo-relevancia (*Pseudo Relevance Feedback*, PRF). Este método consiste en realizar una recuperación normal para obtener un conjunto inicial de documentos relevantes. A partir de ese conjunto se consideran los  $n$  primeros documentos, se seleccionan los  $m$  términos más relevantes de dichos documentos según un esquema de pesado particular y, finalmente, se expande la consulta original con esos términos seleccionados, lanzando una nueva recuperación de documentos. Como se ha explicado anteriormente en la Sección 3, Terrier permite aplicar de forma automática este mecanismo PRF utilizando una técnica de expansión de consulta llamada *Bo1*. La Tabla 1 muestra los resultados obtenidos aplicando esta técnica y utilizando diferentes valores para los parámetros de configuración  $n$  y  $m$ .

Como se puede observar en la Tabla 1, los resultados obtenidos son muy similares a pesar de modificar los parámetros de configuración  $n$  y  $m$ . El mejor valor MAP se consigue utilizando los tres primeros documentos y los 15 o 20 mejores términos para expandir la consulta.

##### 4.2. Expansión MeSH

El siguiente grupo de experimentos consistió en expandir la consulta utilizando términos de la ontología médica MeSH. Se propusieron dos tipos de expansión:

- *MeSH-ALL*, en la que se consideraron

Expansión	P	R	MAP
MeSH-ALL	0,059	0,470	0,191
MeSH-ACE	0,066	0,526	0,236

Tabla 2: Resultados de la expansión utilizando la ontología médica MeSH

todas las categorías MeSH y los campos de registro MH, ENTRY, PRINT ENTRY y DE para expandir la consulta. Además, no se eliminaron posibles términos repetidos durante la expansión.

- **MeSH-ACE**, en la que se consideraron únicamente las categorías A, C y E de MeSH y los campos de registro MH, ENTRY y PRINT ENTRY. Además, sí se eliminaron los términos repetidos durante esta expansión.

Se han seleccionado las categorías A (*Anatomy*), C (*Diseases*) y E (*Analytical, Diagnostic and Therapeutic Techniques, and Equipment*) porque se corresponden mejor con la temática de la colección (Chevallet, Lim, y Radhouani, 2006) y consiguen mejorar el resultado de la expansión como demuestran otros trabajos realizados (Díaz-Galiano et al., 2010; Díaz-Galiano, Martín-Valdivia, y Ureña-López, 2009).

Los resultados obtenidos mediante este tipo de expansión se muestran en la Tabla 2. Como era de esperar, la expansión en la que se utilizaron sólo tres categorías MeSH obtuvo los mejores resultados, con una diferencia de +23,56 % en cuanto a valor MAP. Otra de las razones de este comportamiento fue la eliminación de términos repetidos en la expansión *MeSH-ACE*, ya que con ello se lograba introducir menos ruido en la consulta.

### 4.3. Filtrado de la expansión utilizando ganancia de información

En el siguiente método se propuso filtrar los términos que, a partir de la ontología MeSH, se añadieron a la consulta durante la expansión *MeSH-ALL* explicada en la sección anterior. Por tanto, el objetivo en este grupo de experimentos fue eliminar aquellos términos que podían no beneficiar el proceso de recuperación de información. Para ello, se tomó el corpus completo y se calculó el valor de ganancia de información (*Information Gain*, IG) de cada palabra.

La ganancia de información es una medida

basada en la entropía de un sistema, es decir, en el grado de desorden de un sistema (Shannon, 1948). Esta medida nos indica cuánto se reduce la entropía de todo el sistema si conocemos el valor de un atributo determinado. De esta forma, podemos conocer cómo se relaciona el sistema completo con respecto a un atributo, o lo que es lo mismo, cuánta información aporta dicho atributo al sistema. La IG para un término  $x$  se calculó mediante la siguiente fórmula:

$$IG(x) = \log(n) + \frac{n_x^3}{n^2} \log \frac{n_x}{n} + \frac{n_x^3}{n^2} \log \frac{n_x}{n} \quad (2)$$

donde:

- $n$  es el número total de documentos en la colección
- $n_x$  es el número total de documentos que contienen la palabra  $x$

Hay que tener en cuenta que un término MeSH puede componerse de varias palabras, por lo que el peso del término será la suma de los pesos de sus palabras, en este caso la suma del valor IG de cada palabra. Para los experimentos relacionados con IG se propusieron distintas estrategias de filtrado que permitían una configuración en base a tres parámetros:

- *Fórmula de pesado utilizada para cada palabra*
- *Ítems sobre los que se calcula el peso (palabras o lemas)*
- *Selección de los mejores*, para expandir con aquellos términos que tuvieran un mayor peso. A su vez, esta selección podía estar basada en un porcentaje o en un número determinado (umbral)

En base a estos parámetros se evaluaron las siguientes configuraciones o estrategias utilizando como fórmula de pesado la ganancia de información:

- **TPIG**. Los ítems que se consideraron fueron palabras o términos (T) y la selección o filtrado de los mejores se hizo por porcentaje
- **LPIG**. Se consideraron los lemas (L) y se filtró por porcentaje
- **TNIG**. Se tuvieron en cuenta las palabras y se filtraron aquellos términos con

Filtrado	Umbral	<i>P</i>	<i>R</i>	MAP
<b>LNIG</b>	$n = 1$	0,063	0,493	<b>0,228</b>
	$n = 2$	0,063	0,488	0,212
	$n = 3$	0,062	0,486	0,210
	$n = 4$	0,062	0,485	0,205
	$n = 5$	0,062	0,484	0,205
	$n = 10$	0,062	0,483	0,205
	$n = 15$	0,062	0,482	0,205
	$n = 20$	0,062	0,482	0,205
<b>LPIG</b>	10 %	0,062	0,485	0,210
	20 %	0,062	0,485	0,206
	30 %	0,062	0,484	0,206
	40 %	0,062	0,483	0,205
	50 %	0,062	0,482	0,205
<b>TNIG</b>	$n = 1$	<b>0,064</b>	<b>0,495</b>	0,223
	$n = 2$	0,062	0,485	0,208
	$n = 3$	0,062	0,486	0,202
	$n = 4$	0,062	0,487	0,199
	$n = 5$	0,062	0,486	0,199
	$n = 10$	0,062	0,484	0,195
	$n = 15$	0,062	0,486	0,197
	$n = 20$	0,062	0,487	0,196
<b>TPIG</b>	10 %	0,062	0,486	0,198
	20 %	0,063	0,488	0,198
	30 %	0,062	0,487	0,198
	40 %	0,062	0,487	0,197
	50 %	0,062	0,487	0,197

Tabla 3: Resultados del filtrado de la expansión utilizando Ganancia de Información (IG)

un peso mayor a un umbral  $n$  preestablecido

- **LNIG.** Se consideraron los lemas y se expandió con aquellos términos que tenían un peso mayor a un umbral  $n$  preestablecido

Por último, los umbrales de porcentaje analizados han sido 10, 20, 30, 40 y 50, mientras que el número de términos con mayor peso añadidos ha sido estudiado sobre los valores  $n = \{1, 2, 3, 4, 5, 10, 15, 20\}$ . La Tabla 3 muestra los resultados obtenidos con este tipo de filtrado para las distintas configuraciones contempladas.

Como se puede observar en los resultados obtenidos en la Tabla 3, la configuración que consigue el valor más alto de MAP es la que combina lemas y un determinado número de términos a añadir (LNIG). De hecho, el valor más alto de MAP (0,228) se obtiene seleccionando únicamente el lema con mayor peso ( $n = 1$ ), y progresivamente decrece el valor de MAP a medida que añadimos más términos

a la expansión. En cuanto a la diferencia de considerar lemas o las propias palabras como ítem sobre el que calcular el peso, el utilizar lemas se comporta mejor, aunque la diferencia no es tan importante (+2,24 % y +6 % con umbrales  $n = 1$  y 10 %, respectivamente).

#### 4.4. Filtrado de la expansión utilizando información mutua

Además de la ganancia de información propusimos utilizar otra fórmula de pesado diferente para las palabras: la Información Mutua (IM). La información mutua (*Pointwise Mutual Information*, PMI) es una medida de asociación o relación entre palabras utilizada a menudo en áreas como la teoría de la información y la estadística (Cover y Thomas, 2006). La información mutua se calcula mediante la siguiente fórmula:

$$PMI(x, y) = \log_2 \frac{m \cdot n_{xy}}{n_x n_y} \quad (3)$$

donde:

- $n$  es el número total de documentos en la colección
- $n_x$  es el número total de documentos que contienen la palabra  $x$
- $n_y$  es el número total de documentos que contienen la palabra  $y$
- $n_{xy}$  es el número total de documentos que contienen ambas palabras  $x$  e  $y$
- $m$  es el número total de palabras diferentes en la colección

Siguiendo el mismo tipo de configuraciones explicadas en la sección anterior para la ganancia de información, aplicamos diferentes valores y porcentajes de umbral, considerando tanto lemas como las propias palabras a la hora de expandir. Los resultados obtenidos utilizando información mutua como fórmula de pesado se muestran en la Tabla 4.

Como se puede observar en la Tabla 4, el comportamiento del sistema es similar al obtenido utilizando ganancia de información como fórmula de pesado. La configuración LNMI (lemas y valores específicos de umbral) es la que alcanza el mejor resultado MAP, también con 0,228 para el valor de umbral  $n = 1$ . Otra característica de estos resultados es que, al igual que antes, no existen importantes diferencias entre las distintas configuraciones, destacando la igualdad de valores

Filtrado	Umbral	<i>P</i>	<i>R</i>	MAP
<b>LNMI</b>	$n = 1$	0,063	0,493	<b>0,228</b>
	$n = 2$	0,063	0,488	0,212
	$n = 3$	0,062	0,486	0,210
	$n = 4$	0,062	0,485	0,205
	$n = 5$	0,062	0,484	0,205
	$n = 10$	0,062	0,483	0,205
	$n = 15$	0,062	0,482	0,205
	$n = 20$	0,062	0,482	0,205
<b>LPMI</b>	10 %	0,061	0,477	0,200
	20 %	0,061	0,476	0,200
	30 %	0,061	0,476	0,200
	40 %	0,061	0,476	0,200
	50 %	0,061	0,476	0,200
<b>TNMI</b>	$n = 1$	<b>0,064</b>	<b>0,495</b>	0,223
	$n = 2$	0,062	0,485	0,208
	$n = 3$	0,062	0,486	0,202
	$n = 4$	0,062	0,487	0,199
	$n = 5$	0,062	0,486	0,199
	$n = 10$	0,062	0,484	0,195
	$n = 15$	0,062	0,486	0,197
	$n = 20$	0,062	0,487	0,196
<b>TPMI</b>	10 %	0,062	0,480	0,200
	20 %	0,062	0,479	0,200
	30 %	0,062	0,480	0,200
	40 %	0,062	0,480	0,200
	50 %	0,062	0,480	0,200

Tabla 4: Resultados del filtrado de la expansión utilizando Información Mutua (MI)

MAP obtenidos cuando se utilizan porcentajes de umbral tanto para lemas como para palabras (LPMI y TPMI).

### 5. Análisis global de resultados

En esta sección se trata de comparar los resultados obtenidos durante toda la experimentación llevada a cabo en este trabajo. Para ello, se han escogido los mejores resultados MAP de cada experimento mostrado anteriormente, además del caso base, que consistió en recuperar información con las consultas originales sin aplicar ningún tipo de expansión. Todos estos resultados se muestran en la Tabla 5.

Como se puede observar, el experimento en el que se utilizó expansión PRF con Terrier combinando los tres primeros documentos recuperados y los 15 mejores términos de dichos documentos obtuvo el mejor resultado (0,247). La mejora con respecto al resultado obtenido cuando no se aplicó expansión de consulta (caso base) no fue muy relevante (+5,56 %). En cambio, los experimentos

Experimento	MAP
Caso base (consultas sin expansión)	0,234
Expansión PRF ( $n = 3, m = 15$ )	<b>0,247</b>
MeSH-ALL	0,191
MeSH-ACE	0,236
LNIG ( $n = 1$ )	0,228
LNMI ( $n = 1$ )	0,228

Tabla 5: Resumen de los mejores resultados obtenidos

que se llevaron a cabo utilizando la ontología MeSH como recurso de conocimiento para expandir las consultas introdujeron demasiado ruido en el proceso de recuperación y sólo con la expansión *MeSH-ACE*, en la que se redujo el número de categorías MeSH utilizadas, se consiguió mejorar levemente el caso base (+0,85 %). Como cabía esperar, la expansión de la consulta con todos los términos MeSH relacionados (*MeSH-ALL*) obtuvo el peor resultado comparado con el caso base (-22,51 %), ya que no se aplicó ningún tipo de filtrado sobre los términos que se añadieron. Por último, los experimentos en los que sí se aplicó un filtrado sobre los términos a expandir, relacionados con ganancia de información e información mutua, obtuvieron resultados similares, no mejorando tampoco el resultado alcanzado con el caso base (-2,63 %).

### 6. Conclusiones y trabajo futuro

En este trabajo se ha llevado a cabo una serie de experimentos sobre la expansión de consulta para colecciones médicas utilizando la ontología médica MeSH. El marco de evaluación utilizado ha sido ImageCLEFmed, un track perteneciente a las conferencias CLEF que se encarga de evaluar sistemas de recuperación de imágenes médicas que contienen texto asociado. En concreto, además de realizar una expansión PRF automática mediante el motor de recuperación utilizado, se han realizado varios experimentos expandiendo la consulta original con términos de la ontología MeSH sin ningún tipo de filtrado, escogiendo previamente determinadas categorías para expandir o aplicando un filtrado previo según dos funciones de pesado para cada término: ganancia de información e información mutua.

Como era de esperar, la aplicación de una expansión PRF en este tipo de sistemas mejora levemente los resultados obtenidos cuando no se aplica ningún tipo de expansión, como

ocurre en la mayoría de sistemas RI tradicionales. Por otro lado, la conveniencia de la expansión de las consultas con términos MeSH es difícil de determinar. La disparidad en los resultados obtenidos con dicha técnica dificulta llegar a la conclusión de cuándo es más conveniente aplicar la expansión. En ocasiones, la consulta expandida mejoró claramente a la original, mientras que en otras expansiones se empeoraron los resultados obtenidos. No obstante, se puede deducir que la expansión de la consulta con términos procedentes de una ontología médica como MeSH introdujo demasiado ruido durante el proceso de recuperación, al menos mediante las técnicas aplicadas en este trabajo. Además, los métodos de filtrado descritos utilizando ganancia de información e información mutua adolecieron de asignar más peso a aquellos términos que se componían de un mayor número de palabras, por lo que una mejora para estos métodos consistiría en trabajar con alguna reducción de este valor, como la media, el mínimo o el máximo.

### Bibliografía

- Aronson, A.R. y T.C. Rindflesch. 1997. Query expansion using the UMLS metathesaurus. En D.R. Masys, editor, *Proceedings of the 1997 AMIA Annual Fall Symposium*, páginas 485–489.
- Chevallet, Jean-Pierre, Joo-Hwee Lim, y Saïd Radhouani. 2006. A structured visual learning approach mixed with ontology dimensions for medical queries. *Accessing Multilingual Information Repositories. Lecture Notes in Computer Science*, páginas 642–651.
- Cover, Thomas M. y Joy A. Thomas. 2006. *Elements of information theory (2. ed.)*. Wiley.
- Díaz-Galiano, M. C., M. T. Martín-Valdivia, y L. A. Ureña-López. 2009. Query expansion with a medical ontology to improve a multimodal information retrieval system. *Computers in Biology and Medicine*, 39(4):396–403.
- Díaz-Galiano, Manuel Carlos. 2011. *Recuperación de información multimodal basada en integración de conocimiento*. Ph.D. tesis, Universidad de Jaén.
- Díaz-Galiano, M.C., M.A. García-Cumbreras, M.T. Martín-Valdivia, y A. Montejo-Ráez, 2010. *Knowledge Integration using Textual Information for Improving ImageCLEF Collections*, volumen 32 de *The Information Retrieval Series*, páginas 295–313. Springer Berlin Heidelberg.
- Gobeill, J., D. Theodoro, E. Patsche, y P. Ruch. 2009. Taking benefit of query and document expansion using mesh descriptors in medical imageclef 2009. *Working Notes of CLEF*.
- Hersh, W., S. Price, y L. Donohoe. 2000. Assessing thesaurus-based query expansion using the umls metathesaurus. *Proc AMIA Symp*, páginas 344–348.
- Hersh, William R., Henning Müller, y Jayashree Kalpathy-Cramer. 2009. The imageclefmed medical image retrieval task test collection. *Journal of Digital Imaging*, 22(6):648–655.
- Karamanis, Nikiforos. 2007. Text mining for biology and biomedicine. *Computational Linguistics*, 33(1):135–140.
- Lana-Serrano, S., J. Villena-Román, y J.C. González-Cristóbal. 2008. Miracle at image-clefmed 2008: Evaluating strategies for automatic topic expansion. En *Working Notes of the 2008 CLEF Workshop, Aarhus, Denmark*.
- Müller, Henning, Thomas Deselaers, Thomas Martin Deserno, Paul Clough, Eugene Kim, y William R. Hersh. 2006. Overview of the ImageCLEFmed 2006 Medical Retrieval and Medical Annotation Tasks. En *CLEF*, volumen 4730 de *Lecture Notes in Computer Science*, páginas 595–608. Springer.
- Nilsson, Kristina, Hans Hjelm, y Henrik Oxhammar. 2005. SUiS - cross-language ontology-driven information retrieval in a controlled domain. En Stefan Werner, editor, *Proceedings of the 15th NODALIDA conference*.
- Porter, M.F. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.
- Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1.
- Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27:379–423.