Procesamiento del Lenguaje Natural, Revista nº45, septiembre 2010, pp 21-29 recibido 02-05-10 revisado 17-05-10 aceptado 20-05-10

Un método eficaz de indexación para la recuperación de imágenes en archivos en formato *pdf*[*]

An efficient method of indexing for image retrieval from pdf files

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: Una de las áreas que más interés está despertando actualmente entre los investigadores y usuarios de sistemas de Recuperación de Información es la recuperación de documentos que contengan imágenes relevantes a una necesidad de información. En este caso, el principal objetivo no es la recuperación de los documentos relevantes a la necesidad de información del usuario sino la obtención de las imágenes relevantes a dicha necesidad. En la actualidad, las colecciones de documentos se pueden encontrar en diversos formatos (*html*, *xml*, *pdf*, etc.). En este artículo presentamos un método eficaz para indexar una colección de documentos en formato *pdf* para mejorar la recuperación de imágenes contenidas en los documentos. Los experimentos realizados prueban que el método presentado obtiene mejores resultados que si se realizara una indexación del texto completo.
Palabras clave: Recuperación de información, recuperación de imágenes

Abstract: One of the areas which is presently awakening more interest among researchers and users of Information Retrieval systems is the retrieval of documents containing images which are relevant to a need for information. In this case, the main objective is not the retrieval of the documents relevant to the user's need for information, but the achievement of the images relevant to that need for information. At present, document collections can be found in a variety of formats (*html*, *xml*, *pdf*, etc). In this paper we present an efficient method to index a collection of documents in *pdf* format to improve the retrieval of images contained in documents. The experiments we carried out prove that the method presented here achieves better results than indexing the full text.
Keywords: Information retrieval, image retrieval

1 Introducción

Los sistemas de recuperación de información (SRI) son, actualmente, las herramientas más utilizadas para obtener los documentos de una colección relevantes a una determinada necesidad de información. Tanto en el campo de la investigación como en el comercial, se está prestando enorme atención y realizando grandes esfuerzos para mejorar diferentes aspectos de los SRI: precisión de los resultados, presentación de los documentos o interacción con los usuarios. El funcionamiento general de estos sistemas consiste en presentar todos los documentos, ordenados por el grado de relevancia, como respuesta a una necesidad de información. Una vez obtenida la lista de documentos, el usuario podrá consultarlos para saber si se ajustan a su necesidad y realizará las acciones que considere oportunas. Algunos sistemas facilitan esta tarea ofreciendo un resumen o un extracto de cada documento.

Sin embargo, en numerosas ocasiones, el objetivo de la búsqueda no es la obtención de

[*] 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)"

los documentos que traten sobre las palabras utilizadas en la consulta sino la recuperación de imágenes relevantes a la necesidad de información junto con el documento asociado. Por ejemplo, en el ámbito médico, ante la necesidad de información "*fractura de cúbito*", un especialista podría estar interesado en recuperar, únicamente, los documentos en los que aparezcan imágenes que muestren o estén relacionadas con las fracturas de este hueso. En el ámbito de la arqueología, donde hemos realizado la experimentación que presentamos en este trabajo, el experto desea buscar "*vasijas fenicias*". En este caso, el usuario no está interesado en los textos que versen sobre este tipo de recipiente sino en aquellos documentos que contengan fotografías o dibujos que muestren vasijas fenicias.

2 Trabajos relacionados

En la actualidad existen diversos sistemas de recuperación de imágenes y sus documentos asociados. La mayoría de ellos están especialmente diseñados y optimizados para trabajar en el dominio biomédico. En (Xu, McCusker y Krauthammer, 2008), los autores proponen un sistema de recuperación de imágenes biomédicas y sus artículos (*papers*) asociados, denominado *Yale Image Finder* (*YIF*). La búsqueda se basa en la consulta de las leyendas de las imágenes, los resúmenes (*abstracts*) o los títulos de los artículos. Además, mediante el uso de un Reconocedor Óptico de Caracteres (OCR), realizan una indexación del texto contenido en las imágenes. Este sistema muestra imágenes de otros documentos relacionadas con la imagen recuperada, permitiendo encontrar documentos relacionados a partir de una imagen de interés. Actualmente, YIF tiene indexadas 513.993 imágenes de libre acceso incluidas en artículos de revistas de *PubMed Central*.

Por su parte, *ARRS Goldminer* (Kahn Jr. y Thao, 2007) proporciona acceso instantáneo a 240.117 imágenes radiológicas publicadas en 262 revistas. La principal novedad es la indexación por conceptos. Utilizan recursos de la Biblioteca Nacional de Medicina de Estados Unidos, principalmente la parte correspondiente a *NIH* (*National Institutes of Health*) para obtener conceptos médicos desde las leyendas de las figuras. Para ello, también hacen uso de la ontología *MeSH* (*Medical Subject Heading*). Como dicen los propios autores, los resultados de la búsqueda en *GoldMiner* dependen de la presencia de palabras específicas en las leyendas de las figuras.

BioText (Hearst et al., 2007) proporciona a los especialistas en medicina nuevas formas de acceder a la literatura científica. Para realizar las búsquedas, *BioText* indexa los resúmenes y las leyendas de las figuras, permitiendo acceder al documento completo a través de las imágenes devueltas. En 2007 el sistema indexó todos los documentos de acceso libre disponible en *PubMed Central*. Posteriormente, algunos de los autores de este trabajo (Divoli, Wooldridge y Hearst, 2010) hicieron un estudio con 20 usuarios especialistas en biomedicina para conocer el alcance y las opiniones sobre la interfaz propuesta para *BioText*.

En (Yu y Lee, 2006), los autores proponen algoritmos para relacionar frases de los resúmenes de los artículos con el contenido de las leyendas de las figuras. En este trabajo presentan una interfaz llamada *BioEx* que muestra un conjunto de imágenes debajo de cada resumen.

Más recientemente, en (Yu et al., 2009) se solicitó a distintos especialistas del dominio biomédico que determinaran cuánta información importante acerca de las figuras se encontraba en las propias leyendas de las figuras frente al resumen y al texto completo del artículo. El estudio concluyó que el uso de las leyendas, los títulos y los resúmenes conlleva una menor comprensión que el uso del texto completo.

En cuanto a la recuperación de imágenes a partir de documentos *pdf*, los autores proponen en (Christiansen, Lee y Chang, 2007) un método para recuperar documentos relevantes a partir de las leyendas de las figuras. Para ello proponen un método para asociar cada leyenda con su figura correspondiente. Sin embargo, utilizan únicamente las leyendas para construir el índice y asumen que todas las leyendas están correctamente situadas debajo de la figura correspondiente.

Como se puede apreciar, los sistemas descritos anteriormente son capaces de obtener 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.

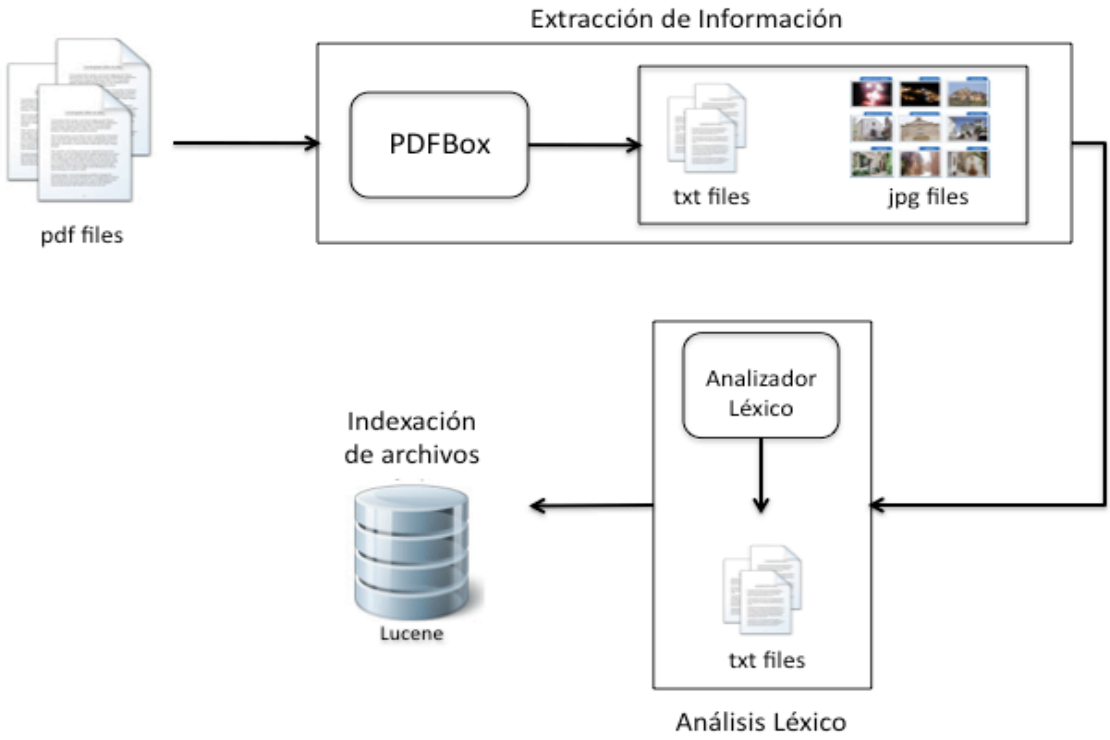

Figura 1: Arquitectura del sistema

En este trabajo presentamos un método eficaz para la indexación de documentos que se encuentran en formato *pdf* y cuyo principal objetivo es la recuperación de las imágenes contenidas en ellos. Aunque en este artículo nos centramos en este tipo de documentos, la técnica presentada es directamente adaptable a cualquier formato en el que se encuentren las colecciones de documentos.

El resto del artículo se estructura de la siguiente forma. En la sección 3 se describe la problemática que pueden tener las distintas propuestas de indexación planteadas en los trabajos relacionados. En la sección 4 se detalla la arquitectura de nuestro sistema. Posteriormente se presenta la colección de documentos utilizada para el desarrollo y experimentación en la sección 5. En la sección 6 se define el procedimiento de enlazado de las imágenes con el texto asociado. En la sección 7 se muestran y discuten los resultados obtenidos en los distintos experimentos y, finalmente, las conclusiones y trabajos futuros se detallan en la sección 8.

3 Indexación para la recuperación de imágenes

Para la indexación, los principales sistemas de recuperación de imágenes basados en el contenido textual utilizan, mayoritariamente, el texto de los títulos, las leyendas de las figuras y los resúmenes.

Los títulos y los resúmenes suelen describir, de forma precisa, el contenido del documento. Sin embargo, el texto contenido en estas secciones no es suficientemente significativo para recuperar, con precisión, las imágenes relevantes a una consulta.

Por otro lado, la indexación del texto completo, aunque obtiene buena cobertura, provoca una disminución considerable en la precisión. Esto es debido a que se recuperan todos los documentos relevantes a las consultas junto con las imágenes incluidas en ellos. Por ejemplo, la consulta "*estructuras fenicias*" devolvería todas las imágenes de los documentos que contienen esas palabras. Se recuperarían, por tanto, multitud de imágenes no relevantes simplemente por el simple hecho de pertenecer a documentos relevantes.

La indexación del texto contenido en las leyendas de las figuras es, con toda probabilidad, la estrategia más habitual para recuperar imágenes contenidas en documentos. Las palabras de las leyendas suelen describir, con cierta exactitud, la imagen que representan. Además, con este enfoque se recuperan únicamente las imágenes relevantes y no todas las contenidas en el documento, aumentando considerablemente la precisión. Los motores de búsquedas descritos en la sección anterior han sido especialmente diseñados y optimizados para recuperar imágenes de documentos del dominio biomédico. En este tipo de documentos, las leyendas de las imágenes suelen ser bastante descriptivas (a veces, una única leyenda contiene el mismo número de palabras que el resumen del artículo). Pero, desafortunadamente, no siempre se dispone de leyendas tan narrativas. En muchas ocasiones, únicamente aparecen frases como "*Figura 6*", "*Cuadro IV*" o "*Lámina III*". En estos casos, el sistema es incapaz de recuperar las imágenes consultando las leyendas.

Nuestra propuesta se fundamenta, por tanto, en la indexación, junto con las leyendas de las imágenes, de las secciones del documento que hacen referencia a imágenes. Por ejemplo, si en el texto aparece el párrafo "*... todavía se conservan restos de ajuares funerarios que pueden ser datados con anterioridad al s. III a.C., tal como los que aparecen en la fotografía 3, en el que se puede ver un anillo tallado con una deidad griega…*", se utilizará, para construir el índice, una ventana de palabras anterior y posterior a la expresión "*fotografía 3*", de forma que se almacenen términos como "*ajuares*", "*funerarios*", "*anillo*", "*s. III a.C.*", "*deidad*" o "*griega*". Estos términos del índice estarán asociados a la imagen correspondiente a la *Fotografía 3*.

4 Arquitectura del sistema

El sistema consta de tres módulos bien diferenciados. En la primera etapa se extrae la información contenida en los documentos para que puedan ser manipulada, posteriormente, con mayor facilidad. El segundo módulo es el encargado de realizar el análisis léxico del texto extraído de los documentos para localizar las secciones que, en una etapa posterior, serán las que se utilicen para construir el índice. Por último, en el tercer módulo se lleva a cabo la indexación de los términos seleccionados en el análisis anterior.

En la Figura 1 se muestra, de forma esquemática, la interconexión de cada uno de los módulos.

4.1 Extracción de información

El objetivo de este módulo es extraer, por un lado, el texto del documento y, por otro, las imágenes contenidas en él. Para el tratamiento de los archivos *pdf* se ha hecho uso de la herramienta *PDFBox*[1]. Esta librería Java de código abierto permite trabajar con documentos en formato *pdf*. Entre sus funciones se encuentra la de extraer el contenido en este tipo de documentos.

La principal razón de extraer el texto del documento y almacenarlo en un fichero de texto plano es para que el análisis léxico que se realiza en el módulo siguiente se lleve a cabo de una forma eficiente.

4.2 Análisis léxico

En este módulo se realiza un análisis léxico de los ficheros de texto generados en el módulo anterior para identificar las palabras o términos que hacen referencia a imágenes. En este trabajo, a estas palabras o términos los denominaremos TRI (*Término Representativo de Imagen*). Haciendo un estudio de los documentos y con la ayuda de un experto del dominio, para este trabajo se diseñó un analizador léxico capaz de reconocer los siguientes TRI: *Figura*, *Lámina*, *Imagen*, *Gráfico*, *Fotografía*, además de posibles abreviaturas como *Lam.* o *Fig.*

Cabe resaltar la posibilidad de añadir nuevos TRI al analizador léxico. De esa forma, el sistema es flexible y adaptable a cualquier dominio y a cualquier idioma. Si quisiéramos trabajar con una colección de documentos escritos en inglés, bastaría con añadir nuevos TRI como *Figure*, *Image* o *Picture*, entre otros.

En este módulo, además, se genera el texto que será indexado posteriormente. Para ello, por cada TRI distinto identificado, se crea un nuevo fichero de texto plano. En este fichero se almacena el propio TRI junto con los caracteres que se encuentran en las posiciones anteriores y posteriores formando una ventana.

[1] *The Apache Software Foundation*. *Apache – PDFBox*, http://pdfbox.apache.org/

Para que el sistema pueda funcionar correctamente, cada fichero estará relacionado con la imagen a la que hace referencia y con el documento de la colección original al que pertenece. Los TRI de un mismo documento que hacen referencia a la misma imagen se almacenan en el mismo fichero.

Por ejemplo, si en un documento aparecen los fragmentos "*...helicoidal (u.e. 4) con olambrilla y una orla transversal (fig. II), ocupando una amplia extensión del sondeo…*" y "*FIG. 2. Planta perteneciente al suelo u.e. 4, en el Corte B.*", el contenido del fichero a indexar correspondiente al TRI "Figura 2" contendrá el siguiente texto "*con olambrilla y una orla transversal (fig. II), ocupando una amplia extensión. FIG. 2. Planta perteneciente al suelo u.e. 4, en el Corte B.*".

Estos ficheros, tal como aparece en la Figura 1, conforman la entrada del tercer módulo, es decir, contienen a los términos que se indexan y sobre los que se realizan las búsquedas.

Lo más destacable de esta estrategia es que, en realidad, lo que se hace es una expansión de la indexación de las leyendas de las imágenes con el contexto donde se encuentra la descripción de las propias imágenes.

4.3 Indexación de archivos

El objetivo de este módulo es construir un índice con el contenido de los ficheros generados en el módulo anterior. Para la creación de este índice se ha utilizado *Lucene* (Cutting et al., 2008). *Lucene* es una librería que implementa todas las características de un motor de recuperación de información. Está implementada en *Java* y es de código abierto.

Tanto para la indexación como para la búsqueda se ha utilizado un analizador en castellano para realizar un preprocesado consistente en la eliminación de las palabras vacías y reducción a la raíz (*stemming*).

5 Colección de documentos

La colección de documentos seleccionada para llevar a cabo la evaluación del sistema pertenece al ámbito de la arqueología. Concretamente se trata de las memorias anuales redactadas a partir de diferentes intervenciones realizadas en provincias de la comunidad andaluza.

En estas memorias existe un considerable número de imágenes. Los profesionales de la arqueología están interesados en recuperar documentos que contengan imágenes relevantes a una necesidad de información. De ahí el interés de disponer de una aplicación como la que se presenta y evalúa en este artículo.

El principal inconveniente de las colecciones de documentos que existen en la red (véase *PubMed Central*) es que no disponen de un conjunto de pruebas con el que poder evaluar las técnicas desarrolladas.

Para poder evaluar correctamente el sistema, como se verá en la sección 6, hemos contado con la estimable ayuda de un experto en el dominio, que se ha encargado de etiquetar las imágenes pertenecientes a los documentos de la colección que son relevantes a un conjunto de necesidades de información. De esta forma se ha podido calcular la precisión, la cobertura y la medida F.

Uno de los principales inconvenientes de esta colección es la ausencia de leyendas en muchas de las imágenes.

El número total de documentos de la colección es de 80. En la Tabla 1 se muestran algunos datos estadísticos de la colección.

	Nº total	**Media por doc.**
TRI	1365	17.06
Imágenes	657	8.21
Páginas	607	7.58

Tabla 1: Estadísticas generales de la colección

6 Enlazado de las imágenes con el texto asociado

Una de los procedimientos de mayor complejidad de este sistema es la asociación o enlazado de las imágenes de cada documento con su correspondiente fichero de texto indexado por *Lucene*.

Como se describió en la sección 3.1, para la extracción del texto y de las imágenes de los documentos se hizo uso de la herramienta *PDFBox*. El principal problema es que la extracción de las imágenes se realiza según el orden en que éstas aparecen en el documento mientras que los TRI pueden encontrarse en diferentes posiciones en el texto del documento. Además, un TRI se repite cada vez que hace referencia a una misma imagen y, por supuesto, éstos no tienen ninguna relación con la posición de la imagen.

En la Figura 2 se muestra un ejemplo de un extracto de documento de la colección. En la sección sombreada aparece el TRI "*Lam: III*", que hace referencia a la segunda imagen, en secuencia, de la página. Para que el proceso de indexación sea correcto, es necesario disponer de un procedimiento capaz de asociar el fichero formado por la ventana del TRI "*Lam: III*" con su imagen correspondiente.

En esta intervención arqueológica de urgencia se han seguido, las siguientes fases:
1ª Fase.- Estudio bibliográfico tanto de las fuentes literarias como de los informes y memorias de las intervenciones arqueológicas llevadas a cabo con anterioridad en éste área.
2ª Fase.- A través de la excavación arqueológica se ha confirmado la ausencia de materiales y estructuras de entidad en este solar, si bien el análisis de los escasos restos de estructuras, así como del proceso deposicional nos ha proporcionado una visión de la evolución de dicho solar a lo largo del tiempo.
Los pasos seguidos han sido los siguientes:
A/ Limpieza del área de excavación.
B/ Si bien en el proyecto se planteó la excavación hasta una cota de -4,50 metros siendo esta la máxima que se va a rebajar para la construcción del ascensor, la aparición de los niveles vírgenes en una cota superior, no hizo necesario alcanzar dicha cota.
C/ Se ha intervenido un único área de excavación de 4 x 3 m., que si bien en el proyecto se planteó en una zona que afecta a la próxima construcción de la bodega y cercana a la localización del ascensor, por causas de seguridad nos centramos en el solar alejándonos de la medianera más al Oeste del mismo, desplazándonos 1m hacia el Este, ya que según información del arquitecto de la obra es la que se encuentra en peor estado. Este área se ha dividido en Unidades de Registro de 1 Metro x 1 Metro numeradas del 1 al 12. El Norte arqueológico se situó en la misma dirección que el norte geográfico.
En cuanto a la cota cero, se situó en un punto del interior del solar a una cota de 8.340 s.n.m. Esta cota se arrastró desde un punto situado, según el mapa topográfico de Andalucía, escala 1:10.000, en la calle Feria.
D/ Realización de la planimetría del área excavada y su correspondiente reportaje fotográfico.
E/ Una vez concluida la intervención propiamente dicha, se efectuará el seguimiento de la obra en el resto del área de remoción de tierras que no se ha excavado en la presente intervención.
Una vez realizada la cuadriculación del área en Unidades de Registro, la excavación se realizó siguiendo el llamado "Método Harris".
3ª Fase.- Elaboración de informe memoria de esta intervención en el cual se trataron los siguientes puntos.
A/ Análisis de las posibles estructuras existentes.
B/ Levantamiento de la planimetría de las estructuras existentes.
C/ Estudio de los materiales.

VII. DESARROLLO GENERAL DE LA INTERVENCIÓN.

Las unidades estratigráficas detectadas durante la presente intervención reflejan la evolución deposicional en el área excavada, mostrando los distintos periodos históricos que han afectado a la zona.
En primer lugar es destacable la ausencia de Unidades estratigráficas construidas o estructuras de importancia habiendo sido de interés el análisis de los procesos deposicionales en el solar en cuestión.
Las primeras estructuras aparecidas se relacionan con la edificación precedente y parcialmente derribada para la ejecución del proyecto actual. Dichas estructuras consistían en la cimentación de un muro, el depósito de nivelación del suelo del patio que ocupaba el lugar de la presente intervención (consistente en escombros compactados con restos materiales contemporáneos) un pozo ciego y varias atarjeas e intrusiones recientes de tuberías y bajantes introducidos en reformas efectuadas en la construcción mencionada (Lam: II y V). La aparición de un pozo ciego y una atarjea invade el perfil Sur del área, superponiéndose a un pozo de agua de cronología mudéjar (Lam: V y VI).
La única estructura destacable en la presente excavación arqueológica, se encuentra a una cota muy alta, y se trata de un suelo de ladrillos a sardinel que aparece roto por atarjeas y tuberías contemporáneas (Lam: III). Este suelo presenta un acusado buzamiento hacia el Norte, como reflejan las cotas tomadas, oscilando las superiores entre los 8.045 s.n.m. la máxima y 7.915 s.n.m. la superior mínima. Se trata de un suelo de patio con una distribución radial. La que sería su parte central en origen, que es la que presenta una cota mas baja, vendría relacionada con un pozo ciego que serviría de desagüe del patio (Lam: IV). El escaso material aparecido en sus depósitos relacionados nos da una cronología para esta estructura del siglo XVII.
Continuando la secuencia estratigráfica, se pueden destacar una serie de depósitos de cronología mudéjar.
La única estructura de esta cronología mudéjar la constituye un pozo de agua que aparece junto al perfil Sur y que, a una cota superior que oscila entre los 6.971 s.n.m. y los 6.925 s.n.m., correspondería a mediados del siglo XIV (Lam: V, VI y VII).

LÁM. II. Cimentaciones y atarjeas de la casa precedente

LÁM. III. Suelo a sardinel roto al sur por las cimentaciones contemporáneas.

485

Figura 2: Ejemplo de documento con un TRI que hace referencia a una imagen

Para resolver este problema se ha desarrollado un procedimiento que realiza el proceso de enlazado. Este procedimiento se basa, principalmente, en la nomenclatura utilizada para identificar los documentos que se genera en el módulo de *Análisis Léxico* y que se indexan en el sistema. Haciendo uso del número asociado a cada TRI ("*figura 4*", "*imagen 7*", etc.), se establece la asociación del fichero que contiene el TRI con su fichero de imagen correspondiente generado en el módulo de *Extracción de la Información*.

De esta forma, cada fichero de texto que se indexa y que contiene las secciones de texto correspondiente al mismo TRI, estará asociado con un fichero de imagen y, por tanto, los términos del índice harán referencia a una imagen concreta.

7 Resultados y discusión

7.1 Preparación del entorno experimental

Tal como se ha comentado en la sección 4, para evaluar el sistema contamos con la colaboración de un experto del dominio que valoró la relevancia de las imágenes respecto a cinco necesidades de información. De esta forma, pudimos construir *Gold Standard* con todas las imágenes de la colección.

Con estas premisas, se consideró hacer uso de la *precisión*, *cobertura* como métricas para evaluar la eficacia del sistema de recuperación de imágenes.

- *Precisión*: proporción de imágenes relevantes recuperadas por el sistema que son relevantes para el experto.
- *Cobertura*: proporción de las imágenes relevantes que han sido identificadas por el sistema.

$$precisión(P) = \frac{\#imágenes\_relevantes\_recuperadas}{\#imágenes\_recuparadas} \quad (1)$$

$$cobertura(C) = \frac{\#imágenes\_relevantes\_recuperadas}{\#imágenes\_relevantes} \quad (2)$$

Con objeto de tener un único valor que permita comparar los distintos experimentos, se utilizó la *medida-F* (3) de (Van Rijsbergen, 1979). En nuestro caso hemos utilizado $\beta = 1$ de forma que la precisión y la cobertura reciben la misma relevancia en el cálculo.

$$F = \frac{(\beta^2 + 1) * precision * cobertura}{\beta^2 * precision + cobertura} \quad (3)$$

Con idea de probar el sistema con diferentes ejemplos y tipos de consultas, se le solicitó al experto que definiera cinco necesidades de información. Para resolver dichas necesidades, el experto decidió utilizar cuatro consultas de un único término y una consulta formada por dos términos. En la Tabla 2 se muestran las

necesidades de información junto con las palabras utilizadas para realizar la consulta.

Necesidad de información	Consulta
Imágenes que contengan las estructuras que aparecen en las excavaciones	estructuras
Imágenes de estructuras funerarias	tumbas
Imágenes de planos de excavaciones	planos
Imágenes con algún tipo de material encontrado en las excavaciones	materiales
Imágenes en las que se aprecien estructuras de época romana	estructuras AND romanas

Tabla 2: Necesidades de información utilizadas en las pruebas

Los experimentos se realizaron en base a dos parámetros: utilización o no del preprocesado de reducción a la raíz (*stemming*) y tamaño de la ventana utilizada para construir el texto a indexar, tal como se muestra en la Tabla 3.

Preprocesado	Tamaño de la ventana (en caracteres)
con *stemming*	50
	75
	100
sin *stemming*	50
	75
	100

Tabla 3: Parámetros utilizados en la experimentación

Además, se diseñó un conjunto de prueba para utilizarlo como un punto de partida (*baseline*) y poder comparar los resultados obtenidos por el sistema. La construcción de este *baseline* consistió en la indexación del texto completo de los documentos que forman la colección. En este tipo de indexación, cuando los términos de la consulta se encuentran en un documento, los sistemas de recuperación de imágenes devuelven todas las imágenes contenidas en el documento.

7.2 Resultados y discusión

El objetivo de nuestro sistema es la recuperación, de la forma más precisa, de las imágenes contenidas en una colección de documentos en formato *pdf* y que respondan a determinadas necesidades de información.

Los resultados de la experimentación se han obtenido utilizando 3 tamaños de ventana (50, 75 y 100 caracteres) y haciendo uso o no de reducción a la raíz en la indexación y en las consultas. En la Tabla 4 se muestran los valores de *precisión*, *cobertura* y *medida-F* obtenidos para las 5 necesidades de información propuestas por el experto haciendo uso de *stemming*. Los resultados obtenidos sin hacer *stemming* fueron algo más bajos para todos los tamaños de ventana.

Los resultados obtenidos por nuestro sistema se comparan con el *baseline* descrito en la sección anterior. Como se puede apreciar, los resultados globales de nuestro sistema superan significativamente, en *precisión* y *medida-F,* los del *baseline*.

En cuanto a la *cobertura*, haciendo una indexación del texto completo se obtienen muy buenos resultados (100% para 4 de las 5 consultas) y un valor medio de 0.87. Sin embargo, los valores obtenidos para la precisión son significativamente más altos en nuestro sistema que para el *baseline* (0.79 frente a 0.27). Esto es debido a que, al utilizarse el texto completo para construir el índice, se recuperan todas las imágenes (tanto relevantes como irrelevantes) contenidas en los documentos en los que aparecen las palabras de la consulta.

Respecto al tamaño de la ventana, los mejores resultados globales para la *medida-F* se obtienen con la de 50 caracteres. Únicamente para la consulta "*estructuras*", el valor de esta medida es mayor para la ventana de 100 caracteres. Se puede apreciar cómo varían los valores de la *precisión* y la *cobertura* en función del tamaño de la ventana. A medida que se incrementa el tamaño de la ventana aumenta la *cobertura* pero disminuye la *precisión*. Con esta observación se deduce que, si se desea usar el contexto para la indexación y recuperación de imágenes, el tamaño de las secciones del documento utilizadas para construir el índice debe ser de unos 50 caracteres alrededor del TRI. Se probaron tamaños de ventana menores a 50 pero los resultados obtenidos fueron algo inferiores.

Consulta	*Baseline*			Sistema (50)			Sistema (75)			Sistema (100)		
	P	**C**	**F_1**	**P**	**C**	**F_1**	**P**	**C**	**F_1**	**P**	**C**	**F_1**
estructuras	0.14	**1**	0.24	**0.59**	0.58	0.58	0.54	0.74	0.62	0.52	0.82	**0.64**
tumbas	0.22	**1**	0.37	**0.92**	0.86	**0.89**	0.81	0.96	0.88	0.77	0.96	0.85
planos	0.12	**1**	0.21	**0.82**	0.77	**0.79**	0.73	0.72	0.72	0.68	0.78	0.73
materiales	0.05	**1**	0.1	**0.60**	0.85	**0.7**	0.45	0.82	0.58	0.37	0.88	0.52
estructuras AND romanas	0.85	**0.33**	**0.48**	**1**	0.14	0.25	**1**	0.14	0.25	**1**	0.14	0.25
Valor Medio	0.27	**0.87**	0.42	**0.79**	0.64	**0.7**	0.71	0.68	0.69	0.67	0.72	0.69

Tabla 4: Resultados obtenidos para las cinco necesidades de información utilizando *stemming*

Cabe destacar los valores de *cobertura* obtenidos en todas las pruebas para la consulta número 5. En este tipo de sistemas, el éxito de los resultados depende de la presencia de las palabras de la consulta en las leyendas y en las ventanas del texto indexadas. Para la necesidad de información "*Imágenes en las que se aprecien estructuras de época romana*" se utilizó la expresión "*estructuras romanas*". Los valores bajos de cobertura obtenidos indican que, en los documentos que contienen imágenes relevantes a esta necesidad de información, no aparecen los términos de la consulta. Este hecho se corrobora si observamos que, con la indexación del texto completo, la cobertura es, únicamente, del 33%. Nuestro sistema fue capaz de recuperar 9 de las 66 imágenes relevantes etiquetadas por el experto mientras que, indexando el texto completo, se recuperaron 22 de las 66 imágenes. Esta diferencia de 13 imágenes se debe a que, como se explicó anteriormente, el *baseline* devuelve todas las imágenes del documento. De ahí que, con este método, se obtenga un valor inferior para la precisión (85% frente al 100% que se obtiene en nuestro sistema para cualquier tamaño de ventana).

8 Conclusiones y trabajo futuro

Hemos presentado un método de indexación de documentos en formato *pdf* para la recuperación de imágenes contenidas en dichos documentos.

Los resultados obtenidos demuestran que los sistemas de recuperación de imágenes mejoran considerablemente si, para construir el índice, se utilizan, junto con las leyendas, las secciones del documento que referencian a las imágenes. Esta mejora será más considerable en las colecciones de documentos donde las leyendas apenas ofrecen información descriptiva sobre las imágenes.

En este artículo se ha presentado un primer acercamiento para la mejora de la recuperación de imágenes contenidas en documentos. Como trabajo futuro nos planteamos experimentar con documentos del ámbito biomédico puesto que son los especialistas en biomedicina quienes más demandan este tipo de aplicación. También ampliaremos la experimentación con otros tipos de formato de documentos (XML, HTML, …).

Uno de los aspectos que abordaremos será la mejora de la cobertura mediante la expansión de las consultas. Para ello utilizaremos otras secciones donde haya información relevante sobre las imágenes del documento y haremos uso de recursos semánticos (especialmente cuando experimentemos con documentos en inglés y del dominio biomédico).

Otro aspecto en el que intentaremos realizar mejoras será en la definición de la ventana a utilizar alrededor de un TRI. Si bien con la ventana de 50 caracteres se han obtenido buenos resultados, pensamos que con otros tipos de ventana los resultados podrán mejorarse significativamente. En este sentido, definiremos ventanas que tengan una sintaxis y una semántica correcta como, por ejemplo, utilizando frases que se encuentren entre signos de puntuación.

Referencias bibliográficas

Christiansen, A., D. Lee y Y. Chang. 2007. Finding relevant PDF medical journal articles by the content of their figures. *En Proc. SPIE* Vol. 6516

Cutting, D., M. Busch, D. Cohen, O. Gospodnetic, E. Hatcher, C. Hostetter, G. Ingersoll, M. McCandless, B. Messer, D. Naber y Y. Seeley. 2008. *Apache Lucene*. http://apache.lucene.org.

Divoli, A., Michael A. Wooldridge, Marti A. Hearst. 2010. Full Text and Figure Display Improves Bioscience Literature Search. *PLoS ONE* 5(4): e9619.

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.

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

Van Rijsbergen, CJ. 1979. *Information Retrieval*. Second Edition. Ed. Butterworths. Londres.

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.

Yu, H. y M. Lee. 2006. Accessing bioscience images from abstract sentences. *Bioinformatics* 22(14): e547–56.

Yu, H., S. Agarwal, M. Johnston y A. Cohen. 2009 Are figure legends sufficient? Evaluating the contribution of associated text to biomedical figure comprehension. *J Biomed Discov Collab* 4: 1.