Procesamiento de Lenguaje Natural, Revista nº 48 marzo de 2012, pp 73-80 recibido 16-01-12 revisado 06-03-12 aceptado 18-03-12

Experimentación en la Búsqueda de Imágenes a partir de Características Visuales y Textuales: Fusión Tardía y Expansión de la Consulta

Experimentation in Image Search based on Visual and Textual Features: Late Fusion and Query Expansion

Rubén Granados Muñoz
NLP&IR Research Group
ETSI Informática, UNED, Madrid, Spain
rgranados@lsi.uned.es

Ana García-Serrano
NLP&IR Research Group
ETSI Informática, UNED, Madrid, Spain
agarcia@lsi.uned.es

Noelia Méndez Fernández
ETSI Informática, UNED,
Madrid, Spain

Xaro Benavent García
Dpto de Informática
Universitat de València, Spain
xaro.benavent@uv.es

Resumen: La recuperación de información multimedia es uno de los retos que actualmente se afrontan en el entorno de la web o en grandes colecciones de objetos multimedia (audio, video, imágenes y textos). En este artículo se presenta la experimentación realizada para mejorar la calidad de la búsqueda en una colección de imágenes anotadas como es IAPR TC-12. Se muestra cómo la fusión multimedia mejora los resultados de búsquedas monomodales basadas solo en texto o en imagen, y además se propone una aproximación semántica para el subsistema textual a partir del análisis detallado de las anotaciones de la colección y de la expansión de las consultas, que mejora los resultados de todos los experimentos (automáticos) previos en la tarea de recuperación de imágenes con la misma colección.
Palabras clave: Recuperación de información multimedia, Recuperación basada en contenido, Recuperación de información textual, Fusión multimedia, Expansión de consulta.

Abstract: Multimedia information retrieval is one of the main challenges addressed in the Web or in big multimedia collections (audio, video, image and text). This paper presents an experimentation to improve the search quality in an annotated images collection like IAPR TC-12. It is shown how multimedia fusion improves the monomodal search results based just in text or image, and it is also proposed a semantic approach in the textual subsystem based on a detailed analysis of the annotations of the collection and in the query expansion, which improves previous results of all (automatic) experiments in the image retrieval task.
Keywords: Multimedia information retrieval, Content-based retrieval, Textual information retrieval, Multimedia fusion, Query expansión.

1 Introducción

Este trabajo muestra los experimentos realizados y las mejoras conseguidas en la recuperación de imágenes anotadas gracias a la fusión multimedia, a la incorporación de información semántica en las anotaciones, y a la expansión semántica de la consulta. El escenario de experimentación lo aporta la tarea de recuperación de imágenes fotográficas de ImageCLEF 2008 (Arni et al., 2009) y su colección de evaluación IAPR-TC12 (Grubinger et al., 2006).

Se empieza describiendo la colección utilizada, y se realiza una introducción a la fusión multimedia, contextualizando el trabajo

ISSN 1135-5948 © 2012 Sociedad Española para el Procesamiento de Lenguaje Natural

con el análisis de las aproximaciones seguidas por algunos de los grupos participantes en la tarea correspondiente de ImageCLEF. A continuación se describe el sistema de búsqueda de imágenes junto con los algoritmos de fusión multimedia analizados, y se muestran y analizan los experimentos y resultados obtenidos.

Finalmente se describe la expansión semántica propuesta, a partir del análisis detallado de las anotaciones de la colección y con la incorporación de información semántica para la expansión de la consulta. Se verá cómo las técnicas propuestas de fusión multimedia a nivel de decisiones con esta expansión semántica mejoran los resultados previos.

2 Colección IAPR TC-12

La colección empleada en la experimentación es pública y, además de las imágenes anotadas textualmente, se dispone de un conjunto de consultas y sus correspondientes juicios de relevancia, necesarios para la experimentación y evaluación.

Se trata de una colección con imágenes de contenido genérico y variado. Contiene 20.000 fotografías tomadas en cualquier parte del mundo por miembros de Viventura[1], una agencia de turismo independiente. La selección de fotos se hizo siguiendo parámetros de diversidad temática (panorámicas, actividades sociales, deportes, gente o no, animales, ciudades, etc.) y variedad de diferencias visuales (según su iluminación, ángulo de la toma, primeros planos, etc.).

Cada imagen tiene una anotación semiestructurada (en inglés, español y alemán), creada por una única persona, que contiene siete campos: identificador, título, descripción textual del contenido, información adicional, proveedor, localización y fecha.

Las consultas (denominadas topics en el foro ImageCLEF) propuestas para la evaluación proporcionan 2 campos textuales, uno con una descripción corta de lo que se busca (*title*) y otro (*narr*) con una descripción más detallada. Además se incluyen 3 imágenes de ejemplo para la búsqueda visual (Figura 1).

Con los juicios de relevancia proporcionados es posible evaluar los experimentos realizados y compararlos con otros similares como por ejemplo los resultados del resto de grupos participantes en la tarea de ImageCLEFphoto en 2008.

<**title**> church with more than two towers </**title**>
<**narr**> Relevant images will show a church, cathedral or a mosque with three or more towers. Churches with only one or two towers are not relevant. Buildings that are not churches, cathedrals or mosques are not relevant even if they have more than two towers. </**narr**>
<**image**> SampleImages/02/16432.jpg </**image**>
<**image**> SampleImages/02/37395.jpg </**image**>
<**image**> SampleImages/02/40498.jpg </**image**>

Figura 1: Ejemplo de *topic* o consulta

3 La fusión multimedia

El objetivo del procesamiento de recursos multimedia es reducir el "gap semántico" que existe entre las caracterizaciones de sus contenidos y los conceptos que aparecen representados.

Las caracterizaciones de los contenidos multimedia se obtienen seleccionando un conjunto de características, que pueden clasificarse según su nivel de abstracción (Attrey et al., 2010): características de bajo nivel (color, textura o forma en el caso de imágenes), obtenidas mediante el procesamiento basado en el contenido, o características de alto nivel, donde se incluyen representaciones semánticas.

El objetivo de la fusión multimedia es el de combinar múltiples modalidades para mejorar los procesos de acceso a la información (recuperación y búsqueda). Es necesario tener en cuenta:

1. Niveles de fusión. Decidir entre fusión a nivel de características (early fusion), fusión a nivel de decisión (decisión-level o late fusión), o fusión híbrida (propone una combinación de ambas, fusionando algunos modos a nivel de características y posteriormente sus resultados a nivel de decisión, con el objetivo de aprovechar las ventajas de cada estrategia).
2. ¿Cómo fusionar? Elegir entre diferentes técnicas.
3. ¿Cuándo fusionar? Las cuestiones relacionadas con la sincronización entre las distintas modalidades o con el tiempo de procesamiento son determinantes a la hora de plantear la estrategia de fusión.
4. ¿Qué fusionar? Para la obtención de información complementaria y no contradictoria.

[1] www.viventura.net

La fusión a nivel de decisión presenta muchas ventajas con respecto a la fusión basada en características. Por ejemplo, mientras que a nivel de características estas pueden tener diferentes representaciones en función del medio, a nivel de decisión habitualmente se utiliza una misma representación. Además, la fusión a nivel de decisión facilita la escalabilidad, al permitir la inclusión o exclusión de nuevos medios en el proceso de fusión.

Como desventaja, está la de no utilizar una correlación explícita entre los distintos medios, y que el uso de múltiples técnicas para obtener cada una de las decisiones locales hace que el proceso de fusión sea complejo y pueda tener un alto coste computacional.

Para evaluar la propuesta de este trabajo sobre fusión tardía, se utiliza la colección IAPR TC-12, y los resultados de la tarea de recuperación de imágenes fotográficas ImageCLEFphoto, que se detalla en el apartado siguiente (y en la Tabla 3).

3.1 Contextualización

La mayoría de los 24 grupos participantes (Arni et al., 2009), con un total de 1.042 experimentos, se basaron en texto (TBIR) y en la combinación de estos con resultados visuales (CBIR) haciendo uso de técnicas de fusión a nivel de decisión (late fusion). La parte textual suele estar basada en diferentes métodos de pesado (BM25, DFR, LM, VSM), con o sin expansión de la consulta (LCA, PRF, tesauros, CFS, Wordnet).

Los resultados obtenidos por las aproximaciones consideradas más relevantes en relación a la propuesta en este artículo, incluyendo los del grupo ganador, se muestran en la Tabla 3, junto con los resultados de nuestro segundo grupo de experimentos.

El mejor resultado de la competición lo obtiene el grupo XRCE (Ah-Pine et al., 2008), que utiliza como recurso externo un tesauro para enriquecer las anotaciones. Como motor de búsqueda usa la herramienta LEMUR, y sigue una aproximación de fusión multimedia basada en "relevance feedback" entre modos.

El grupo DCU (O'Hare, 2008) utiliza el campo *location* para expandir de forma automática con información geográfica. La fusión se lleva a cabo mediante una combinación lineal (0.7 para el texto; 0.3 para la parte visual) y normalizando con MinMax.

IPAL (Gao et al., 2008) desarrolló varios sistemas CBIR y TBIR, a partir de un diccionario léxico y entrenando un modelo de lenguaje. Utiliza cross-media pseudo-relevance feedback. Para fusionar las decisiones se combinan con un método lineal con coeficientes iguales para cada modo.

El grupo INAOE (Jair et al., 2008) realiza su mejor fusión con un peso de 0.2 para el texto y 0.8 para la parte visual (resultados de imagen solo, o imagen y texto).

El grupo NTU (Chang y Chen, 2008) hace expansión de consulta, *relevance feedback* y fusión, normalizando y combinando con igual peso los resultados monomodales.

Para expandir la consulta el grupo CLAC (Demerdash et al., 2008) usa diferentes métodos como pasar los mejores resultados de cada motor de búsqueda (visual y textual) al otro, añadir sinónimos en base a WordNet, o *pseudo-relevance feedback*. Para combinar los resultados, toman como máximo las 3 imágenes mejor colocadas en la lista visual.

CUT (Wilhelm et al., 2008) utiliza Lucene y expande las consultas con un tesauro basado en OpenOffice que deteriora la precisión. La fusión con visual mejora un 37%.

En MMIS (Overell et al., 2008) combinan anotación de imágenes con un filtro con información geográfica. Usan Lucene, y fusionan filtrando para eliminar imágenes no relevantes y combinado rankings con distintos pesos, llegando a la conclusión de que el texto es dominante para esta tarea y colección.

4 Sistema de recuperación multimedia

La arquitectura del sistema propuesto y desarrollado se muestra en la Figura 2. Está formado por 3 subsistemas principales: 1) recuperación basada en características textuales (TBIR); 2) recuperación basada en características visuales (CBIR); y 3) fusión multimedia, que combina los resultados obtenidos en los módulos anteriores mediante la utilización de diferentes algoritmos de fusión a nivel de decisiones (late fusion).

TBIR. Maneja la información textual de las anotaciones (García-Serrano et al., 2008). Este subsistema, basado en la herramienta IDRA (Granados et al., 2009), extrae la información de los distintos campos XML de las anotaciones y realiza el preprocesamiento, consistente en eliminar los caracteres especiales, los signos de puntuación, y las

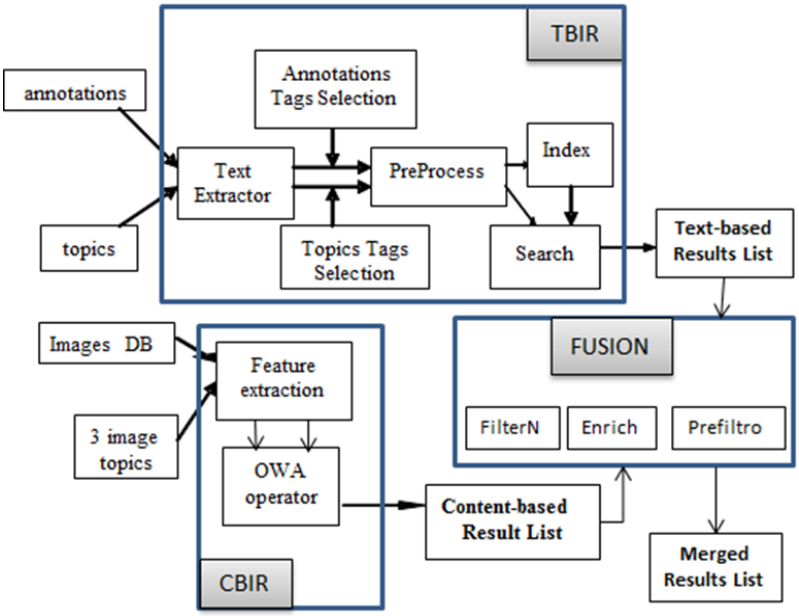

Figura 2. Arquitectura del sistema del primer grupo de experimentos

stopwords (palabras sin contenido). El texto resultante para cada imagen es indexado mediante una aproximación basada en el VSM (Vector Space Model), usando como función de pesado TF-IDF. La función de similitud utilizada entre el vector que representa la consulta y el que representa a la imagen será el coseno, que calcula la relevancia de los resultados con respecto a la consulta. Las consultas (campos *title* y *narr*) serán preprocesadas lingüísticamente de igual manera, y lanzadas contra la indexación para recuperar las imágenes relevantes.

CBIR. Este subsistema se encarga de procesar las características visuales de las imágenes, como se describe con detalle en (Benavent et al., 2010). Las imágenes se representan mediante un vector de características de bajo nivel que describen el color y la textura. Para calcular su similitud con las imágenes de consulta se utiliza la medida de Mahalanobis, que obtiene mejores resultados que la Euclídea. Cada consulta contiene 3 imágenes de ejemplo, teniendo así tres distancias de similitud, que se combinarán en una sola para reordenar todas las imágenes. Esta combinación se hará mediante el operador matemático de agregación OWA con distintos pesos. OWA transforma un número finito de entradas en una única salida. No se asocia un peso concreto con ninguna de las entradas, sino que la magnitud relativa de cada entrada decide qué peso le corresponde a cada entrada. El *orness* indica el grado con el que la agregación se parece a la operación OR. Los mejores resultados se obtienen con un orness(W)=0.3.

Fusión Multimedia. Subsistema encargado de combinar los resultados obtenidos desde TBIR y CBIR, con fusión a nivel de decisión o tardía (late fusion). Se proponen, implementan y evalúan diferentes técnicas y algoritmos:

- FilterN. Elimina de la lista de resultados textuales aquellas imágenes que no aparezcan entre las N primeras de la lista visual.

- Enrich. Enriquece las decisiones tomadas desde un modo con las del otro, recalculando los valores de similitud para una determinada consulta a partir de los resultados obtenidos independientemente desde cada modo:

$$newRel = mainRel + \frac{supRel}{(posRel + 1)} \qquad (1)$$

donde el nuevo valor de similitud (newRel) se calcula en función del valor en el modo textual (mainRel), de la similitud en el modo visual (supRel), y de la posición en la lista de resultados visuales (posRel).

- Prefiltro. Como paso previo a la recuperación basada en el contenido visual, el módulo TBIR decide un conjunto de imágenes que considera no relevantes. El módulo visual trabajará solo con las imágenes restantes. Este prefiltrado se basa en la idea de que el módulo textual captura mejor inicialmente el significado conceptual de una consulta.

5 Primer grupo de experimentos y resultados

Entre todos los experimentos realizados con el sistema descrito, se muestran en la Tabla 1 los mejores TBIR (baseline textual), CBIR (mejor experimento visual), y FUSION (FilterN / Enrich / Prefiltro).

Los resultados se evalúan en términos de MAP y P20. La medida de evaluación MAP (Mean Average Precision) calcula la media sobre todas las consultas de la precisión promedio (AP, Average Precision) de cada una de ellas. AP tiene en cuenta, además del número de imágenes relevantes recuperadas, el orden en que estas se recuperan. La medida P20 (Precisión a 20 o P@20) calcula la precisión teniendo en cuenta únicamente los 20 primeros resultados recuperados, esto es, el número de imágenes relevantes recuperadas entre los primeros 20 resultados, dividido entre 20.

Experimento	Modo	P20	MAP
TBIR	TXT	0.2846	0.2253
CBIR	IMG	0.0679	0.0213
FilterN	TXTIMG	**0.3179**	0.1936
Enrich	TXTIMG	0.3090	**0.2401**
Prefiltro	TXTIMG	0.1385	0.0656

Tabla 1. Resultados iniciales

Destaca la diferencia entre los resultados obtenidos por el módulo TBIR y CBIR, lo que indica la importancia de disponer de anotaciones asociadas a las imágenes.

En cuanto a los beneficios de la fusión multimedia, se observa que en todos los casos se superan los resultados visuales, y que los resultados de texto también son superados cuando se fusionan ambos modos con los algoritmos FilterN y Enrich. Con FilterN (N=10000) se mejoran los resultados de precisión a bajos niveles, aquellos que mostraría un buscador convencional en su primera página, esto es, los únicos que consultaría la gran mayoría de usuarios. El algoritmo Enrich, aparte de mejorar esos valores de precisión a bajos niveles, es capaz también de superar el valor de MAP con respecto al experimento TBIR.

Con esto se muestra la mejora proporcionada por cada una de las técnicas de fusión multimodal, confirmando la complementariedad entre ambos modos y la mejora de los resultados obtenidos.

6 Expansión semántica

En esta sección se propone una modificación del módulo TBIR de recuperación basada en texto mediante la incorporación de información semántica, aunque los beneficios se muestran con una búsqueda basada en Lucene. Se realiza un segundo grupo de experimentos, que mejorarán todos los resultados de lo experimentos participantes en la competición, tanto en la modalidad textual como en la de fusión.

6.1 Análisis de la colección

Examinando lingüísticamente uno por uno los distintos campos presentes en las anotaciones de las imágenes, se observa lo siguiente:

<title>. Puede incluir nombres propios, frases nominales generales, una combinación de ambos, o frases cortas. La longitud media del título en inglés es de 5.35 palabras, el más largo contiene 17 palabras y el más corto 1.

<description>. Contiene una descripción semántica de los contenidos de la imagen con frases cortas y nominales. La longitud media de las descripciones en inglés es de 23.06 palabras, la más larga contiene 85 y la más corta 2. Después de realizar un estudio detallado, se concluye que estas descripciones cumplen un patrón de prioridad: las primeras frases describen la información semántica, y las siguientes los elementos adicionales o información del entorno. La mayoría de las frases cumplen uno de los patrones que se muestran en la Tabla 2. Los símbolos utilizados son: S (sujetos, con o sin adjetivos), V (verbos, con o sin adjetivos), O (objetos, con o sin adjetivos), PA (adjuntos de lugar con preposición de lugar), TA (adjuntos de tiempo con preposición de tiempo) y P (cualquier patrón o combinación de patrones).

Patrón P	Ejemplo
S	A red rose
S-V	A boy is singing
S-TA	A boy at night
S-PA-TA	A boy in a garden at night
S-V-TA	A boy is singing at night
S-V-PA	A boy is singing in a garden
S-V-PA-TA	A boy is singing in a garden at night
S-V-O	A girl is kissing a boy

S-V-O-TA	A girl is kissing a boy at night
S-V-O-PA	A girl is kissing a boy in a garden
S-V-O-PA-TA	A girl is kissing a boy in a garden at night

Tabla 2. Patrones de las descripciones

<notes>. Contiene información adicional en texto libre sobre las imágenes, que no se puede ver directamente o que requiere conocimientos adicionales.

<location>. Contiene el lugar en el que se tomó la imagen, y se divide en dos partes. Primero se muestra la localización exacta y a continuación el país al que pertenece la localización. Algunas imágenes (2.35%) solo contienen información sobre el país, y en esos casos la localización exacta dentro del país no se puede verificar.

6.2 Expansión de la consulta

Tras el análisis anterior se incorpora a la colección información adicional formada, por un lado, por el conjunto de entidades nombradas que aparecen en el texto, identificadas con el NER de C&C[2], y por otro, un subconjunto de conceptos de la ontología MediaMill 101 (Snoek et al., 2006), que están potencialmente presentes en las imágenes de la colección. De los 101 conceptos presentes en la ontología, se hace uso únicamente de los que resultan relevantes para la colección (un total de 12): *animal, bad weather, body of water, building, female, football, people, racing, sport, vehicle, vehicle* and *waterbody*. Para cada uno de ellos, se genera manualmente un archivo de texto con una lista de términos relacionados que se utilizará tanto para anotar los ficheros de anotaciones de las imágenes como para expandir las consultas (Méndez, 2011).

Con esta información, el sistema realizará un paso de anotación previo a la indexación, llevado a cabo en dos fases:

1) Identificación de las entidades nombradas presentes en los campos <title> y <description>, que se almacenarán en un nuevo campo <entities>. El campo <location>, por definición, siempre contendrá entidades, y se tendrán siempre en cuenta.

2) Identificación de los conceptos extraídos de la ontología presentes en los campos <title>, <description> y <notes>, que se añadirán en un nuevo campo <concepts>. Para identificar los conceptos se hará uso de las listas de términos relacionados generadas.

A los campos *title* y *narr* de las consultas se les aplica el mismo preprocesamiento lingüístico y reconocimiento de entidades y conceptos. Además, se realiza (manualmente) la expansión de las consultas con información geográfica relacionada con los nombres de continentes que aparezcan en ella, añadiendo el nombre de los países del continente identificado (en trabajos futuros se refinará esta extensión). Un ejemplo de las posibles mejoras aplicando esta expansión geográfica puede verse con la consulta 23, descrita en el último párrafo de esta sección.

Además, se normalizarán los verbos, se eliminarán los elementos no relevantes (como "etc."), se seleccionarán únicamente frases positivas (eliminando las que contengan "not relevant"), y se extenderán las consultas con los significados de las siglas y con las diferentes formas de escribir una entidad.

La siguiente gramática aplicada al texto de la consulta en el campo <title> permite determinar la forma en que se construye la consulta final:

TOPIC ← OFQUERY | INQUERY | OUTSIDEQUERY | FROMQUERY | WITHQUERY | BASICQUERY
OFQUERY ← (OF_IZDA) of OF_DCHA
INQUERY ← IN_IZDA in IN_DCHA
OUTSIDEQUERY ← OS_IZDA outside OS_DCHA
FROMQUERY ← F_IZDA from F_DCHA
WITHQUERY ← W_IZDA with W_DCHA
BASICQUERY ← (NOT_QUERY) *TEXTO*
OF_IZDA ← exterior view | indoor | view | *TEXTO*
OF_DCHA ← *TEXTO*
IN_IZDA ← *TEXTO*
IN_DCHA ← NE_QUERY *TEXTO*
OS_IZDA ← *TEXTO*
OS_DCHA ← NE_QUERY
F_IZDA ← *TEXTO*
F_DCHA ← NE_QUERY *TEXTO*
W_IZDA ← *TEXTO*
W_DCHA ← NE_QUERY *TEXTO*
NE_QUERY ← *NAMEDENTITY TEXTO*
NOT_QUERY ← *TEXTO* not *TEXTO*

La consulta final estará formada por las tres partes (entidades, conceptos y contenidos), y se podrá añadir una cuarta subconsulta en función de la gramática:

- BASICQUERY: cuando la consulta básica contiene la palabra "not", se crea una

[2] http://svn.ask.it.usyd.edu.au/trac/candc/wiki/NER

consulta que se añade a la original con el operador "NOT".

- OUTSIDEQUERY: se crea una consulta negativa con las entidades que contiene la parte derecha de la consulta (es decir, el texto posterior a "*outside*"), y se añade al resto de la consulta con el operador "NOT".

- INQUERY: en este tipo de consultas se tiene en cuenta el contenido de la narrativa completa, utilizándolo para construir la consulta en el caso de que no se encuentren resultados al utilizar el contenido reducido.

Por ejemplo, la consulta 18 ("sport stadium outside Australia") se identifica como OUSIDEQUERY y, por tanto, la entidad "Australia" en la parte derecha de la consulta se utilizará para crear una consulta con el operador NOT, evitando de este modo recuperar imágenes asociadas con Australia.

Un ejemplo de las mejoras obtenidas gracias a la expansión semántica de la consulta, es la número 23, en la que se menciona uno de los estados de los EEUU (California), y se expande con ciudades de dicho estado, y se consigue pasar de un MAP desde 0 hasta 0.4520.

6.3 Segundo grupo de experimentos

Se ejecutan nuevos experimentos textuales y multimedia (fusionando con el experimento CBIR original) que mejoran sustancialmente los resultados previos, como se observa en la Tabla 3.

Los experimentos de fusión llevados a cabo, con el experimento CBIR original, son:

- Enrich (TXT, CBIR)
- Enrich (TXT+sem, CBIR)
- Enrich (TXT, Prefiltro)
- Enrich (TXT+sem, Prefiltro)

La Tabla 3 muestra los resultados obtenidos (MAP y P20) por este segundo grupo de experimentos, junto con los de los grupos participantes mencionados en la sección 3.1.

Se observa que la expansión semántica mejora los resultados textuales hasta un MAP de 0.3555, sobrepasando al mejor experimento textual automático presentado en la competición (0.3514). La fusión de este nuevo experimento textual con los resultados visuales iniciales del subsistema CBIR se realiza con el algoritmo de fusión que mejor resultados obtuvo (Enrich) en la primera fase de este estudio. El mejor resultado, tanto en MAP como en P20, se obtiene fusionando con el algoritmo Enrich los resultados textuales obtenidos tras la expansión semántica, con los resultados visuales obtenidos tras el prefiltro textual (Prefiltro). En este caso se consigue llegar a un MAP de 0.4265, que supera también al mejor experimento multimedia presentado a la competición (0.4105).

Experimento	**MODO**	**P20**	**MAP**
TBIR	TXT	0.2397	0.2039
TBIR	TXTsem	0.4564	0.3555
Enrich	TXTIMG	0.4256	0.3072
Enrich	TXTsemIMG	0.5449	0.4165
EnrichPrefiltro	TXTIMG	0.4244	0.3124
EnrichPrefiltro	TXTsemIMG	0.5462	**0.4265**
GRUPO	**MODO**	**P20**	**MAP**
XRCE	TXTIMG	0.5731	0.4105
DCU	TXTIMG	0.4756	0.3510
IPAL	TXTIMG	0.4282	0.3109
INAOE	TXTIMG	0.3910	0.3066
NTU	TXTIMG	0.3769	0.2806
CLAC	TXTIMG	0.3538	0.2673

Tabla 3. Resultados y Comparativa

Actualmente se están realizando estos análisis con otra colección multimedia, la proporcionada en ImageCLEF 2011 con imágenes de Wikipedia (Tsikrika et al., 2011), para confirmar estos resultados. Se puede indicar que los de la primera fase de este estudio ya confirman la mejora con la fusión multimedia utilizando un prefiltro textual para CBIR (Benavent et al., 2010).

7 Conclusiones

Se han descrito los experimentos iniciales monomodales, donde se comprueba el mejor funcionamiento de la recuperación basada en texto que la basada en contenido visual. A continuación se experimenta y prueba que la fusión multimodal en esta colección mejora los resultados monomodales, sobre todo en términos de precisión, pero también el MAP con alguno de los algoritmos de fusión empleados. Es importante resaltar que el hecho de que la fusión mejore en casi todos los casos los valores de precisión en los primeros resultados recuperados significa un gran avance, ya que en general, suelen ser los únicos visitados por la mayoría de usuarios.

Posteriormente, con la extensión propuesta para la recuperación textual, basada en un

análisis detallado de las anotaciones de las imágenes, en la incorporación de información semántica, y en el tratamiento de la consulta según su tipo, se consiguen mejorar no solo los resultados del primer grupo de experimentos de los autores del trabajo, sino los del resto de grupos de la competición para la misma colección, tanto en experimentos textuales, como en los experimentos que utilizan fusión multimedia.

8 Agradecimientos

Este trabajo se ha financiado con los proyectos competitivos: MA2VICMR (S2009/TIC-1542, financiado por la Comunidad de Madrid), BUSCAMEDIA (CEN-20091026, financiado por el Ministerio de Industria), HOLOPEDIA (TIN 2010-21128-C02), financiado por el Ministerio de Ciencia e Innovación) y TEC2009-12980, financiado por el Ministerio de Economía y Competitividad.

Bibliografía

Ah-Pine, J., C. Cifarelli, S. Clinchant, G. Csurka, J.M. Renders. 2008. XRCE's Participation to ImageCLEF 2008. En (Peters, 2008).

Arni, T., Clough, P., Sanderson, M., Grubinger, M. 2009. Overview of the ImageCLEFphoto 2008 Photographic Retrieval Task. LNCS 5706 .

Atrey, P.K., M.A. Hossain, A. El Saddik, M.S. Kankanhalli. 2010. Multimodal Fusion for Multimedia Analysis: A Survey. Multimedia Systems 16: 345-379.

Benavent, J., X. Benavent, E. de Ves, R. Granados, A. García-Serrano. 2010. Experiences at ImageCLEF 2010 using CBIR and TBIR Mixing Inf. Approaches. CLEF Notebook.

Chang, Y., H. Chen. 2008. Increasing Relevance and Diversity in Photo Retrieval by Result Fusion. En (Peters, 2008).

Demerdash, O., L. Kosseim, S. Bergler. CLaC at ImageCLEFPhoto 2008. En (Peters, 2008).

Gao, S., J. Chevallet, J. Lim. 2008. IPAL at CLEF 2008: Mixed-Modality based Image Search, Novelty based Re-ranking and Extended Matching. En (Peters, 2008).

García-Serrano, A., X. Benavent, R. Granados, J. M. Goñi-Menoyo. 2008. Some results using different approaches to merge visual and text-based features in CLEF'08 photo collection. LNCS 5706, Evaluating Systems for Multilingual and Multimodal Information Access. Pp. 568-571. ISSN: 0302-9743.

Granados, R., García-Serrano, Ana., Goñi, J.M. 2009. La herramienta IDRA (Indexing and Retrieving Automatically). XXV Conferencia SEPLN, San Sebastián.

Grubinger, M., P. Clough, H. Müller, T. Deselaers. 2006. The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems. In OntoImage'2006, Genova, Italy.

Hair, H., J. González, C. Hernández, A. López, M. Montes, E. Morales, L. Sucar, L. Villaseñor. TIA-INAOE Participation at ImageCLEF 2008. En (Peters, 2008).

Méndez, N. 2011. Modelo de Búsqueda con Información Semántica: Experimentos sobre la colección IAPR TC-12. Tesis de Master. ETSI Informática, UNED.

O'Hare, N., P. Wilkins, C. Gurrin, E. Newman, G.Jones, A. Smeaton. 2008. DCU at ImageCLEFPhoto 2008. En (Peters, 2008).

Overell, S., A. Llorente, H. Liu, R. Hu, A. Rae, J. Zhu, D. Song, S. Rüger. 2008. MMIS at ImageCLEF 2008: Experiments combining Different Evidence Sources. En (Peters, 2008).

Peters, C. Working Notes for the CLEF 2008 Workshop. 2008. ISBN: 2-912335-43-4.

Snoek, C., M. Worring, J. van Gemert, J. Geusebroek, A. Smeulders. 2006. The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. En Proceedings of ACM Multimedia, pp. 421-430.

Tsikrika, T., A. Popescu, J. Kludas. 2011. Overview of the Wikipedia Image Retrieval Task at ImageCLEF 2011. CLEF (Notebook Papers/Labs/Workshop).

Wilhelm, T., J. Kürsten, M. Eibl. 2008. The Xtrieval Framework: ImageCLEF Photographic Retrieval Task. En (Peters, 2008).