UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA
Escuela Técnica Superior de Ingeniería Informática
Departamento de Lenguajes y Sistemas Informáticos

Tesis Doctoral

Fusión Multimedia Semántica Tardía aplicada a la Recuperación de Información Multimedia

Rubén Granados Muñoz
Ingeniero en Informática

Julio, 2013

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA
Escuela Técnica Superior de Ingeniería Informática
Departamento de Lenguajes y Sistemas Informáticos

Tesis Doctoral

Fusión Multimedia Semántica Tardía aplicada a la Recuperación de Información Multimedia

Rubén Granados Muñoz
Ingeniero en Informática

Directora:
Ana García Serrano

Departamento de Lenguajes y Sistemas Informáticos
ETSI Informática
Universidad Nacional de Educación a Distancia

Julio, 2013

Resumen

El auge alcanzado por los dispositivos multimedia ha generado una cantidad ingente de información (fotografías, vídeos, música, etc.) lo que hace necesario el desarrollo de nuevas aproximaciones para la gestión y la recuperación de información multimedia. Por ejemplo, solo en la red social *Facebook* se añaden cada día 300 millones de fotografías nuevas o en *Instagram* que se han subido unos 5.000 millones de imágenes desde su creación en Octubre de 2010 (según el informe "*Internet 2012 in numbers*" publicado por *Royal Pingdom*[1]).

La principal motivación de este trabajo de tesis doctoral es contribuir a gestionar y recuperar información multimedia desde grandes colecciones o repositorios, permitiendo además que el usuario exprese su necesidad de información utilizando una o varias modalidades de información (texto, audio, imagen). Aunque los objetos multimedia pueden ser muy variados, este trabajo se concentra en un escenario de recuperación multimedia de imágenes anotadas, por lo que se dispondrá tanto de metadatos textuales, como de las correspondientes características visuales de las imágenes (color, forma, textura). Sin embargo, como se mostrará en su momento, esta limitación a imágenes anotadas, no limita la aplicabilidad de las aportaciones de este trabajo a otros objetos multimedia.

La búsqueda y la recuperación de contenido multimedia se han resuelto generalmente con estrategias textuales basadas en las descripciones y metadatos de dichos contenidos. Pero en el mundo real no siempre se dispone de información textual de calidad asociada a los objetos multimedia (o al menos que permita encontrar las respuestas esperadas por el usuario). Por otra parte las aproximaciones basadas en los aspectos visuales de imágenes (vídeo), hasta el momento, no alcanzan unos resultados de suficiente calidad. Una vía de solución a este problema complejo, que es el que se plantea en este trabajo, consiste en utilizar estrategias basadas en la combinación o fusión de la información disponible de las distintas modalidades de los objetos en una colección multimedia.

Es conocido que cualquier concepto se describe mejor cuando se utilizan diferentes fuentes de información (Muller, Clough and Desealaers, 2010), por lo que las estrategias de Fusión Multimedia, tratan de aprovechar la sinergia de la información disponible o generada desde

[1] http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/

las distintas fuentes. Tras una breve introducción a la recuperación de información multimedia, en el apartado del estado del arte, se analizan diferentes aproximaciones de fusión multimedia existentes para combinar el conocimiento procedente desde cada uno de los modos de información.

La evaluación experimental de cualquier contribución a la recuperación de información multimedia es imprescindible para mostrar la validez de la misma, pero por otra parte, es muy difícil de realizar aisladamente. Por ello, desde el inicio de este trabajo de tesis se ha participado anualmente en el foro *ImageCLEF*, con el objetivo de evaluar las distintas aproximaciones de fusión multimedia planteadas a lo largo del trabajo en colecciones facilitadas por el foro. En la memoria se describen detalladamente diferentes colecciones de objetos multimedia y los experimentos realizados con ellas.

La aportación más importante de este trabajo es una propuesta de fusión multimedia asimétrica, basada en la inclusión de una fase inicial de prefiltrado textual de la colección original, apoyada en la mayor carga semántica presente en la información textual en comparación con la visual, seguida de la aplicación de algún algoritmo de fusión tardía o a nivel de decisiones (*late fusion*) de todos los resultados monomodales (la elección del algoritmo depende de las características de la colección y la tarea). A esta propuesta se le ha denominado Fusión Multimedia Semántica Tardía, o LSMF por sus siglas en inglés (*Late Semantic Multimedia Fusion*) (Benavent et al., 2013). Tras la experimentación realizada (Benavent et al., 2010) (Granados et al., 2011) se comprueba que con la aproximación LSMF, se cumplen los objetivos planteados al inicio del trabajo, porque se produce una mejora del rendimiento de las soluciones monomodales, a la vez que se simplifica el proceso de búsqueda visual en colecciones de imágenes anotadas, haciendo escalable la tarea sobre grandes colecciones, como se detallará a lo largo de esta memoria.

Abstract

The peak achieved by multimedia devices has generated a huge quantity of information (photographs, videos, music, etc.) which has made the development of new approximations necessary for the management and recovery of multimedia information. For example, in the *Facebook* social network 300 million new photographs are added every day, or in *Instagram* in which some 5,000 million images have been uploaded since its creation in October 2010 (according to the "*Internet 2012 in numbers*" report published by *Royal Pingdom*[2]).

The main motive of this PhD Thesis is to contribute to the management and recovery of multimedia information from large collections or repositories, which also allows the user to express his or her need for information using one or several types of information (text, audio, images). Although the multimedia objects may be very varied, this work concentrates on a scenario for the multimedia recovery of annotated images, for which both textual metadata and the corresponding visual characteristics of the images (colour, shape, texture, etc.) will be available. However, as will be seen at the time, this limit to logged images does not limit the applicability of the contribution to other multimedia objects.

The search for and retrieval of multimedia content has led fundamentally to textual strategies based on the descriptions and metadata of the said content. But in the real world there is not always textual information available associated to the multimedia objects (or at least that which allows the answers expected by the user to be found). On the other hand, the approximations based on the visual aspects of the images (video) do not produce results of sufficient quality. One way of solving this complex problem, which is presented in this work, consists of using strategies based on the combination or fusion of the information available in the different forms of the objects in a multimedia collection.

It is well known that any concept is described better when different sources of information are used (Muller, Clough and Desealaers, 2010), which is why the Multimedia Fusion strategies aim to take advantage of the synergy of the information available or generated from the different sources. After a brief introduction to the recovery of multimedia information, the

[2] http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/

different existing multimedia merger approximations are analysed to combine the knowledge coming from each of the types of information, in the state of the art section.

The main objective of this PhD thesis is to propose a strategy for multimedia recovery based on the combination of the different types of information present in a multimedia object (image with logged texts, texts containing images, videos, etc.), which improve the performance of the monomodal solutions, allows the costly search process of a visual search to be simplified and makes the task of multimedia recovery in large collections scalable.

The experimental evaluation of any contribution to the recovery of multimedia information is essential in order to show its validity, but on the other hand it is difficult to do in an isolated way. That is why, from the beginning of this PhD thesis work there has been annual participation in the *ImageCLEF* forum, with the objective of evaluating the different multimedia fusion approximations presented in collections facilitated by the forum. Different collections of multimedia objects and the experiments carried out on them are described in detail in the report.

The most important contribution of this work is the proposal for asymmetric multimedia fusion, based on the inclusion of an initial phase of the textual prefiltering of the original collection, supported by the semantic included in the textual information in comparison to the visual, followed by the application of a late fusion algorithm or a decision-making level of all of the monomodal results. This proposal has been called LSMF *(Late Semantic Multimedia Fusion)* [Benavent et al 2013]. After the experimentation carried out [Granados et al 2012], [Granados et al 2011] it has been proven that LSMF has reached the expectations for improved performance, at the same time simplifying the multimedia recovery process in collections of logged images, and making the task of multimedia recovery in large collections scalable as will be set out throughout this report.

Índice de Contenidos

Índice de Figuras

Índice de Tablas

Capítulo 1 Introducción

Se describe en este capítulo el escenario donde se enmarca el trabajo desarrollado en esta tesis, así como la motivación que provoca las hipótesis iniciales y los experimentos desarrollados. Se incluyen también los objetivos principales que persigue la tesis y la organización de la memoria.

1.1 Escenario

El auge alcanzado por los dispositivos multimedia en la actualidad ha ido generando día a día una cantidad ingente de información (fotografías, vídeos, música, etc.). Por un lado, surge la necesidad de almacenar toda esta información, ya sea en el ordenador personal de un usuario, en la propia Web (por ejemplo en redes sociales como *Facebook*[3] o *Twitter*[4], o en sitios de compartición de contenidos multimedia como *YouTube*[5] para vídeos, o *Instagram*[6] o *Flickr*[7] para imágenes), o en aplicaciones específicas (por ejemplo sistemas médicos que dispongan de información textual sobre los pacientes, junto con imágenes correspondientes a radiografías u otras pruebas). Por otro lado, aparece el problema de cómo tratar y posteriormente acceder a toda esta información mediante aproximaciones que sean capaces de encontrar, de manera eficiente y precisa, aquella información multimedia que satisfaga las necesidades del usuario.

[3] https://www.facebook.com/
[4] https://twitter.com/
[5] https://www.youtube.com/
[6] http://instagram.com/
[7] http://www.flickr.com/

Es crucial, por lo tanto investigar en la búsqueda o recuperación de información multimedia, a partir de consultas introducidas por el usuario que también podrán tener carácter multimodal[8]: el usuario podrá expresar su necesidad de información utilizando una o varias modalidades de información (texto, audio, imagen). Como se verá más adelante en la revisión del estado del arte, la estrategia básica para abordar la recuperación multimedia es reducirla a la recuperación textual, sobre la base de las anotaciones o metadatos asociados a la información multimedia. Sin embargo, esta solución supone una pérdida irreparable de información variada, que aportan modalidades diferentes a la textual. Y no solo eso, además, la reducción de una imagen a una descripción textual plantea un nuevo aspecto: la no uniformidad de las anotaciones, ya que diferentes anotadores (automáticos o humanos) generarán diferentes anotaciones (incluso no comparables), según su experiencia previa, intención de uso, y otros aspectos.

En cualquiera de los escenarios previamente comentados un objeto multimedia es un vídeo, una imagen, un audio (habla o música), un documento de texto o cualquier combinación de ellos, y pueden hacerse las siguientes afirmaciones:

1. Un objeto de audio (habla), admite una transcripción y puede tener metadatos.
2. Un vídeo puede describirse (con pérdida de información) con un conjunto de imágenes representativas de sus escenas o *clips*, la transcripción del audio asociado y los metadatos, si los hubiera.
3. Una imagen se describe con un conjunto de características textuales y visuales. Entre las características textuales se encuentran: los metadatos técnicos (EXIF, *Exchangeable Image File Format*), una sentencia corta o leyenda y otras descripciones, que pueden contener desde anotaciones realizadas con un vocabulario controlado o social, con conceptos de una ontología, con texto libre con estructura o sin ella (“en primer plano”, “en el fondo”, etc.), con los textos identificados dentro de la imagen (escritos sobre una pancarta, valla publicitaria, etc.), y otras.

[8]Se constata que en la bibliografía se encuentra mayoritariamente “información multimedia” y “consulta multimodal” aunque pueden utilizarse ambos indistintamente (multimedia y multimodal)

Aunque los objetos multimedia puedan ser tan variados, en esta tesis se trabajará en un escenario de recuperación multimedia de imágenes anotadas, por lo que se dispondrá de metadatos textuales y de las correspondientes características visuales de las imágenes.

Cuando en 2008, en el inicio de este trabajo, se estudiaron los escenarios de recuperación de información multimedia, el foro de evaluación de *ImageCLEF* llevaba unos años proporcionando a los investigadores de esta área diferentes colecciones de imágenes anotadas, con sus correspondientes consultas multimodales de evaluación (denominadas *topics*). En aquel momento, los métodos para recuperación se basaban en técnicas textuales o visuales (basadas en el contenido) por separado, con aproximaciones de fusión multimedia que no acababan de aprovechar la sinergia esperada entre ambos modos, como se mostraba experimentalmente. Es el problema abierto que se comenzó a abordar y que será tratado en el trabajo de tesis.

En el apartado referente al estado del arte, se analizan diferentes aproximaciones de fusión multimedia para combinar el conocimiento procedente desde cada uno de los modos de información, y se propone una estrategia de fusión multimedia (contribución principal de la tesis) para mejorar el rendimiento de las aproximaciones monomodales capaz además de simplificar el proceso de recuperación de una manera considerable. En esta área, cualquier contribución debe evaluarse experimentalmente. El marco de evaluación experimental utilizado en este trabajo es el foro internacional *ImageCLEF*, que será descrito detalladamente, así como las colecciones de imágenes empleadas, las tareas abordadas y la metodología de evaluación seguida.

1.2 Motivación

La principal motivación de este trabajo de tesis doctoral surge de la necesidad de gestionar y recuperar información multimedia desde las grandes colecciones existentes hoy en día. Por ejemplo, solo en la red social *Facebook* se añaden cada día 300 millones de fotografías nuevas (según el informe "*Internet 2012 in numbers*" publicado por *Royal Pingdom*[9]). Según el mismo informe, en *Instagram* se han subido unos 5.000 millones de imágenes desde su creación en Octubre de 2010.

[9] http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/

La búsqueda o recuperación de contenido multimedia dentro de una gran colección es una tarea complicada. La solución más utilizada y eficiente hasta hace poco tiempo, es la basada en las descripciones textuales de dichos contenidos, siguiendo las técnicas clásicas de recuperación de información textual. Pero en el mundo real no siempre se dispone de información textual de calidad asociada a los objetos multimedia (o al menos que permita encontrar las respuestas esperadas por el usuario). Cuando no se dispone de este tipo de anotaciones, la única alternativa es realizar búsquedas basadas en el contenido, esto es en las características visuales (de bajo nivel), en el caso de las imágenes. Esta aproximación basada en el contenido (colores, formas, texturas, etc.) ofrece un rendimiento considerablemente menor que la basada en la información textual, debido fundamentalmente al problema del *multimedia semantic gap* o brecha semántica, que hace referencia a la complejidad para entender la información que un usuario percibe a partir de las características de bajo nivel de un objeto multimedia (Smeulders et al., 2000), esto es, la falta de correspondencia entre lo que un sistema puede representar con este tipo de descriptores de bajo nivel y el significado de los objetos multimedia.

La necesidad y la complejidad de la tarea de recuperación de información multimedia, así como el hecho de que sean las aproximaciones basadas únicamente en texto las que mejor han funcionado, motivan, en segundo lugar, el trabajo en esta tesis hacia el análisis de estrategias basadas en la combinación de las distintas modalidades de información disponibles en una colección multimedia en una tarea de recuperación de información. Para ello, se parte de la base de que cualquier concepto es descrito de mejor manera si para ello se utilizan diferentes fuentes de información (Muller, Clough and Desealaers, 2010). De esta idea surge la Fusión Multimedia, para tratar de aprovechar la información disponible o generada desde las distintas modalidades presentes en un objeto multimedia.

Tras la situación identificada de resultados multimedia peores que los monomodales[10] textuales, y el análisis de las técnicas de fusión de información multimedia de la literatura (detallados en el apartado del estado del arte de este documento), se orienta el trabajo de la tesis hacia la propuesta de una estrategia de fusión multimedia que sea capaz de aprovechar las ventajas y particularidades existentes en las informaciones de modo textual y visual de las

[10] Se constata que en la bibliografía se usa mayoritariamente monomodal en vez de monomedia.

imágenes. Para ello, se parte del hecho de que una adecuada combinación entre ambos modos será beneficiosa para mejorar el rendimiento de la recuperación multimedia (Chatzichristofis et al., 2010). Además aunque las técnicas basadas únicamente en el contenido visual de las imágenes ofrecen un rendimiento bastante más bajo que las basadas en texto, debido principalmente al problema del *semantic gap*, se confía en que los métodos de fusión multimedia puedan incrementar la precisión de la recuperación, proporcionando información adicional y complementaria desde las distintas modalidades (Lew et al., 2006).

Analizando los resultados de la tarea de recuperación de imágenes de la colección de Wikipedia propuesta en el foro de evaluación *ImageCLEF*, que ha tenido lugar durante los cuatro primeros años (2008-2011) de este trabajo, puede observarse cómo las aproximaciones textuales resultan imbatibles para las multimodales o para las visuales durante las ediciones de 2008 y 2009 (Tsikrika, 2010), consiguiendo mejores resultados únicamente para algunas consultas. Es en la edición de 2010 cuando por primera vez, y únicamente en el caso de 2 de los grupos participantes (de un total de 11), los mejores sistemas de recuperación basados en fusión multimedia mejoran a sus respectivas aproximaciones textuales (Tsikrika, 2011). De nuevo en la edición del 2011 el mejor sistema está basado en un enfoque multimedia, pero en este caso esto sucede para la mayoría de los grupos participantes (Tsikrika, Popescu and Kludas, 2011).

Como la evaluación experimental de cualquier contribución a la recuperación de información multimedia es imprescindible y muy difícil de realizar aisladamente, desde el inicio de este trabajo de tesis se ha participado anualmente en el foro *ImageCLEF*, con el objetivo de evaluar las distintas aproximaciones de fusión multimedia planteadas a lo largo del tiempo. La contribución más importante de este trabajo es una propuesta de fusión multimedia asimétrica, basada en la inclusión de una fase inicial de prefiltrado textual de la colección original, apoyada en la mayor carga semántica presente en la información textual en comparación con la visual, seguida de la aplicación de algún algoritmo de fusión tardía o a nivel de decisiones (*late fusion*) de los resultados monomodales. A esta propuesta se le denomina Fusión Multimedia Semántica tardía, o LSMF por sus siglas en inglés (*Late Semantic Multimedia Fusion*). Tras la experimentación con esta propuesta, se cumplen las expectativas de mejora del rendimiento, a la vez que se simplifica el proceso de recuperación multimedia, haciendo escalable la tarea sobre grandes colecciones.

1.3 Doctoral Consortium

En Junio de 2012 se celebró en la UNED un *Doctoral Consortium*, organizado por el Departamento de Lenguajes y Sistemas Informáticos de la Escuela Técnica Superior de Informática, proporcionando a investigadores y estudiantes en los campos del Procesamiento del Lenguaje Natural y de la Recuperación de Información una interesante oportunidad para presentar el progreso de sus trabajos y de recibir comentarios y consejos por parte de los investigadores del consorcio, como son Eduard Hovy, Horacio Rodríguez, y Pablo Castells. Los comentarios recibidos ayudaron a completar el trabajo de investigación realizado hasta ese momento y a reforzar esta tesis.

Se enumeran a continuación los principales comentarios/críticas recibidas en el informe del tribunal, en relación al trabajo realizado hasta aquel momento, junto con las acciones llevadas a cabo para tener en consideración dichas sugerencias:

1. *Se destaca la buena definición del trabajo, así como su realización, pero también la posibilidad de que la investigación se quede únicamente en el análisis de los dos tipos principales de fusión (late y early fusion).*

   Aunque la presentación se centró solo en los aspectos principales de la investigación desarrollada, la tesis doctoral presentada incluirá, además del análisis del comportamiento de varios algoritmos de fusión tardía (*late fusion*), la propuesta de una estrategia de combinación multimedia que mejorará los resultados y simplificará considerablemente la complejidad de la tarea de recuperación, haciéndola escalable sobre grandes colecciones multimedia. Esta técnica es descrita en el Capítulo 5, y evaluada en el Capítulo 7.

2. *Se recomienda el estudio de ImageWordNet para completar el estado del arte en cuanto a recursos disponibles*.

   Aunque el estudio de *ImageNet* está fuera del alcance de la tesis, y debido a su intención principal de servir de apoyo a sistemas de recuperación visuales, en el apartado 2.2.3 se incluye una breve descripción de dicho recurso y su posible utilidad dentro de futuros escenarios.

3. *Se propone analizar el tema de la búsqueda de imágenes interactiva, y ver si la interacción del usuario, remarcando de algún modo los aspectos más relevantes de las imágenes de ejemplo de las consultas, o indicando palabras, puede ayudar a mejorar el rendimiento*.

   Se considera que la interacción en la recuperación de imágenes puede estudiarse desde el punto de vista de cómo el sistema ayuda al usuario en la formulación de la consulta, en su traducción, o en la selección de documentos. Se hace un análisis sobre un corpus pequeño (colección Deportes 20, descrita en el apartado 4.2.2), desarrollado por el autor de esta tesis, y se comprueba que una solución automática desarrollada con un conjunto de reglas es mejor que la basada en interacciones con usuarios de tipo general.

4. *Se sugiere la idea de tratar los componentes visuales de las imágenes (color, textura, forma, etc.) por separado, para combinarlos independientemente con la información textual de diferentes maneras, y analizar qué sucede*.

   En base a esta sugerencia se incluye una sección de la tesis (apartados 7.2.2 y 7.3.6) dedicada a evaluar y analizar esta idea, descomponiendo la información visual en grupos de descriptores visuales (color, textura, etc.). La evaluación del rendimiento de esta aproximación se lleva a cabo tanto desde el punto únicamente visual (CBIR), como desde el basado en la estrategia de fusión multimedia propuesta en esta tesis.

5. *Se pide la explicación sobre la adecuación de la estrategia propuesta a la tarea de recuperación de imágenes, diferenciándola de las técnicas típicas de fusión de rankings*.

   Este aspecto se deja claro cuando se describe la propuesta presentada (Capítulo 5) como una aproximación a utilizar entre modos de diferente nivel semántico como son el texto y la imagen. La técnica presentada también podría ser aplicada a recuperación multimedia de vídeos, música o habla, donde los diferentes modos presentes (vídeo y audio) también poseen un nivel de carga semántica menor a la

de la información textual. Se hace una prueba sobre el corpus Deportes20 incluyendo un tercer modo: el audio (habla).

6. *Para dar mayor profundidad al estudio del tema, se sugiere incluir el análisis de técnicas como el aprendizaje automático (Machine Learnig) o la regresión logística*.

   Aunque dentro de la propuesta de fusión multimedia presentada en esta tesis no se aplican técnicas de aprendizaje automático (salvo para la recuperación basada en el contenido visual), en la parte dedicada al estado del arte (Capítulo 3) se mencionan y describen brevemente algunas de las técnicas más utilizadas dentro de las tareas de análisis multimedia analizadas.

7. *Se recomienda analizar el tema de la normalización de las listas de resultados a fusionar y su conveniencia.*

   En el capítulo referente al estado del arte, y concretamente en el apartado 3.6, se lleva a cabo una detallada revisión de las principales técnicas de normalización utilizadas en la literatura. Posteriormente, dentro del capítulo dedicado a la experimentación (apartado 7.3.4) se analiza la influencia de aplicar este tipo de técnicas dentro del esquema de fusión multimedia propuesto.

La mayoría de los comentarios recibidos han sido tenidos en cuenta para mejorar el trabajo de investigación realizado, aunque alguno de ellos tras el intento inicial queda como trabajo futuro.

1.4 Objetivos

El principal objetivo de este trabajo de tesis es proponer una estrategia de recuperación multimedia basada en la combinación de las distintas modalidades de información presentes en un objeto multimedia (imagen con anotaciones textuales, textos conteniendo imágenes, vídeos, etc.), la cual sea capaz de mejorar el rendimiento de las soluciones monomodales, así como de simplificar el proceso de búsqueda visual y hacer escalable la tarea de recuperación multimedia sobre grandes colecciones.

Con este fin, y centrándose en el escenario descrito de recuperación multimedia de imágenes, puede desgranarse el objetivo principal en los siguientes puntos:

1. Revisar las técnicas de recuperación multimedia de imágenes presentes en la literatura. Se describirán tanto los métodos basados en las anotaciones textuales o metadatos asociados a las imágenes (TBIR, *Text Based Image Retrieval*), como aquellos que utilizan el contenido o descripción visual de las mismas (CBIR, *Content Based Image Retrieval*).

2. Estudiar el concepto de fusión multimedia, con especial interés en sus aplicaciones relacionadas con la recuperación multimedia de imágenes.

3. Comprender y describir el problema de la brecha semántica multimedia (*multimedia semantic gap*), con el fin de poder intentar resolverlo o reducir sus efectos.

4. Proponer y evaluar una estrategia de fusión multimedia capaz de aprovechar la complementariedad existente en las distintas modalidades de información presentes en una colección de imágenes para mejorar el rendimiento de la recuperación, en dos direcciones:

    a. mejorar los valores de precisión del conjunto de imágenes recuperadas

    b. simplificar el complejo y costoso proceso de recuperación visual (CBIR) mediante:

        i. la inclusión de una fase de prefiltrado semántico textual, para hacer escalable la tarea de recuperación sobre grandes colecciones multimedia

        ii. el refinamiento de la consulta con ejemplos visuales negativos (contraejemplos)

5. Analizar, describir e implementar distintos algoritmos de fusión tardía o a nivel de decisiones, que serán aplicados y evaluados tanto independientemente como formando parte de la estrategia de fusión propuesta en esta tesis.

6. Estudiar la conveniencia de incluir procesos de normalización de las listas de resultados monomodales previos a la fase de fusión multimedia tardía.

7. Evaluar la conveniencia de enriquecer las anotaciones textuales asociadas a una imagen mediante el uso de recursos externos de conocimiento.

8. Analizar el comportamiento de la fase prefiltrado textual y de la propuesta de fusión multimedia en función de la complejidad y la carga visual de las consultas multimodales, analizando los beneficios de la interacción del usuario para definir/modificar las consultas y sus parámetros.

9. Desarrollar e implementar una herramienta que incluya todas las funcionalidades desarrolladas a lo largo de la elaboración del trabajo incluido en esta tesis (indexación textual, recuperación, visualización de índices, preprocesamientos, algoritmos de fusión, evaluación de resultados, etc.), que permita la construcción de un sistema avanzado de recuperación de información multimedia parametrizable según la colección.

10. Evaluar las distintas aproximaciones desarrolladas en colecciones provenientes de foros internacionales de evaluación, como CLEF, y publicar los resultados relevantes en congresos y revistas nacionales e internacionales de relevancia.

11. Plantear posibles líneas de investigación futura en base a los resultados observados.

Los objetivos planteados serán abordados independientemente y de manera conjunta a lo largo del desarrollo de esta memoria, justificando las aportaciones con experimentos que conducirán finalmente a un conjunto de conclusiones sobre la principal aportación de esta tesis: LSMF (fusión multimedia semántica tardía).

1.5 Organización de la memoria

Se describe a continuación la estructura seguida en esta memoria para el desarrollo de esta tesis:

PARTE 1 (Estado del arte). En esta parte se hace un repaso sobre el estado del arte en el campo de la Recuperación de Información Multimedia, empezando con una descripción

detallada de la Recuperación de Información basada en texto, junto con las técnicas y modelos más conocidos, siguiendo con la justificación de manejar Información Multimedia en sus distintas modalidades, enfocando el problema en la Recuperación Multimedia de Imágenes, para finalmente describir con detalle los aspectos relacionados con sus escenarios de uso y los trabajos más relevantes llevados a cabo en el área. Se compone de tres capítulos:

- Capítulo 2. Recuperación de Información Multimedia. Se enumeran los principales modelos de recuperación existentes, y se repasan las diferentes técnicas de preprocesamiento de información textual.

- Capítulo 3. Fusión Multimedia. Se describe el concepto de fusión multimedia, detallando los distintos niveles y métodos existentes, así como los principales algoritmos de fusión tardía, prestando especial atención a aquellos empleados en la tarea de recuperación multimedia de imágenes.

- Capítulo 4. Marco de Evaluación. Se detalla la metodología de evaluación seguida en los experimentos incluidos en la memoria, y las colecciones utilizadas.

PARTE 2 (Propuesta y Experimentación). En esta segunda parte se presenta una detallada descripción de la propuesta realizada, junto con el conjunto de experimentos llevados a cabo. Inicialmente se explica la idea y motivación de la aproximación propuesta, describiendo los distintos componentes o fases que son necesarios. A continuación se muestra el entorno de recuperación multimedia diseñado e implementado, que se utilizará para ejecutar los experimentos. Finalmente, y tras mostrar los resultados obtenidos en los distintos experimentos desarrollados, se incluyen las conclusiones obtenidas que estarán relacionadas con las aportaciones de este trabajo. Se incluyen tres capítulos:

- Capítulo 5. Propuesta: Fusión Multimedia Semántica Tardía. Incluye la descripción de la estrategia de fusión multimedia principal aportación en esta tesis.

- Capítulo 6. Entorno desarrollado para la Recuperación Multimedia de Imágenes. Se describe el entorno de recuperación multimedia de imágenes desarrollado, que incluye los sistemas monomodales textual (TBIR) y visual (CBIR), y el módulo de fusión.

- Capítulo 7. Experimentación. Incluye la descripción de todos los experimentos llevados a cabo, junto con los resultados obtenidos, y el análisis de los mismos.

PARTE 3 (Conclusiones). Se compone de los siguientes capítulos:

- Capítulo 8. Principales aportaciones. Resume las principales aportaciones obtenidas con el trabajo realizado.

- Capítulo 9. Líneas futuras de trabajo. Se enumeran algunas de las posibles líneas de trabajo futuras en relación al estado de la investigación alcanzado.

- Capítulo 10. Producción científica. Muestra las publicaciones del autor de esta tesis durante el desarrollo de la misma.

PARTE 1: ESTADO DEL ARTE

En esta parte se hace un repaso sobre el estado del arte en el campo de la Recuperación de Información Multimedia, empezando con una descripción detallada de la Recuperación de Información basada en texto, junto con las técnicas y modelos más conocidos, siguiendo con la necesidad de manejar Información Multimedia en sus distintas modalidades, para describir con detalle los aspectos relacionados con la Recuperación Multimedia de Imágenes, sus escenarios de uso y los trabajos más relevantes llevados a cabo en el área.

Capítulo 2 Recuperación de Información Multimedia

La necesidad de manejar la gran cantidad de información multimedia generada hoy en múltiples escenarios, como grandes empresas, colecciones personales, redes sociales, o en todo tipo de aplicaciones (por ejemplo médicas, con informes textuales, imágenes de radiografías, señales de análisis clínicos, etc.), hace imprescindible disponer de estrategias de recuperación de información multimedia eficientes y precisas.

Un sistema de información multimedia tendrá que ser capaz de soportar todo tipo de información de naturaleza muy distinta (texto, imagen, video, audio), y facilitar tareas como su almacenamiento, recuperación, presentación, etc. Este tipo de sistemas conllevan una dificultad mucho mayor que los sistemas basados únicamente en documentos textuales, aunque la forma más básica y tradicional de abordar la tarea de Recuperación de Información Multimedia ha sido desde un punto de vista textual en la mayoría de las herramientas

existentes, utilizándose anotaciones o información en forma de metadatos asociados a imágenes, audio o videos.

El Diccionario de la Lengua de la Real Academia Española define multimedia (del inglés *multimedia*) como adjetivo para lo que utiliza conjunta y simultáneamente diversos medios, como imágenes, sonidos y texto, en la transmisión de una información. Según esta definición, una imagen anotada será un objeto multimedia.

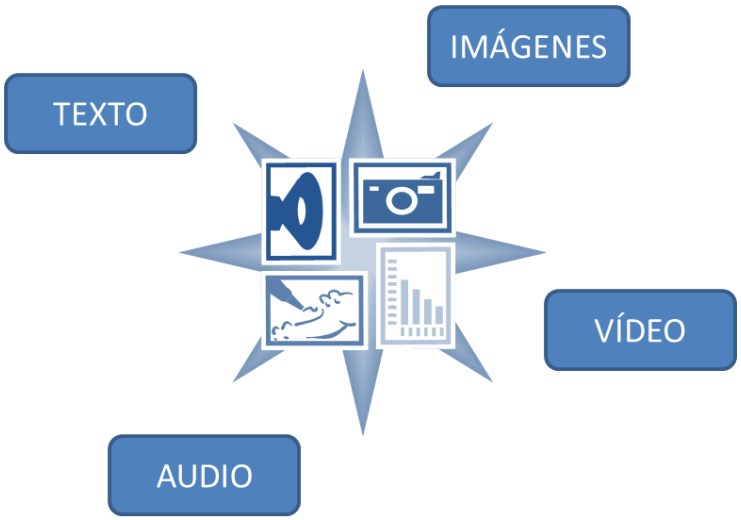

Figura 2.1 Información Multimedia

La información multimedia, junto con los avanzados sistemas y dispositivos existentes en la actualidad, tiene uso y aplicación en una gran variedad de campos, por no decir todos, como pueden ser: el entretenimiento, la educación, el arte, la medicina, la investigación, la publicidad, etc. Por ejemplo en el ámbito de la música, que para algunos es un lenguaje universal, es fácil identificar todos los tipos de información multimedia: el primero, y más evidente, el audio que escuchamos con cada canción; de esas canciones muchas veces nos interesa saber qué dicen, y buscamos las letras, que almacenamos como texto; también cuando de un concierto se toman fotografías para recordar ese momento tendremos información en forma de imágenes; por último, cuando el concierto en el que estuvimos se difunde, hay información en modo vídeo.

Internet y la Web son el más claro ejemplo para observar el crecimiento de información multimedia existente. Cada día se genera más y más cantidad de este tipo de información en páginas de alojamiento y compartición de vídeos (como *Youtube*), de imágenes (como

Instagram, *Picasa*[11] o *Flickr*), o de música (audio). También las redes sociales como *Facebook*, *Twitter* o *Pinterest*[12], generan cada día grandes cantidades de objetos multimedia. Tampoco hay que olvidar las colecciones personales producidas por usuarios individuales gracias al fácil acceso a todo tipo de dispositivos multimedia (cámaras digitales, software de edición de audio, etc.) por parte de un gran número de personas en la actualidad.

2.1 Recuperación de Información

El término *Recuperación de Información* se acuñó en 1952 (Chowdhury, 2010), y fue ganando popularidad en la comunidad científica de 1961 en adelante. El concepto suele ser definido en un sentido muy amplio (van Rijsbergen, 1999). Baeza-Yates (Baeza-Yates and Ribeiro-Neto, 1999) explica la diferencia entre la Recuperación de Información y la Recuperación de Datos, destacando que los datos se pueden organizar en estructuras definidas, como tablas o árboles, para recuperar exactamente lo que se quiere, mientras que el texto no posee una estructura clara y no resulta fácil crearla. Define la Recuperación de Información del siguiente modo: "*dada una necesidad de información y un conjunto de documentos, ordenar los documentos de más a menos relevantes para esa necesidad y presentar un subconjunto de aquellos de mayor relevancia*". Ya en 1986, en (Salton and McGill, 1986) se da una definición parecida: "*selección del subconjunto de documentos adecuados a la necesidad de información de un usuario entre un conjunto más amplio existente en una base de datos documental*".

Según (Martínez Méndez, 2004) se identifican dos grandes etapas cuando se plantea la tarea de la recuperación de información: por un lado, la selección de un modelo que permita calcular la relevancia de un documento frente a una consulta y, por otro, el diseño de algoritmos y estructuras de datos que implementen este modelo de forma eficiente.

Tradicionalmente la Recuperación de Información se ha venido asociando a colecciones de documentos textuales, pero la creciente proliferación de objetos multimedia (imágenes, vídeo, audio) ha hecho necesaria la ampliación de este concepto. El escenario actual exige actualizar las técnicas de recuperación textual, para ser utilizadas sobre colecciones de imágenes, vídeos o sonidos. Es a lo que se llama Recuperación de Información Multimedia.

---

[11] http://picasa.google.com
[12] http://pinterest.com

Conceptualmente el reto no cambia. Se trata de recuperar información, esto es, buscar información relevante para las necesidades del usuario, las cuales tienen cada vez más que ver con objetos multimedia. Los métodos de acceso a esta información dependerán del tipo: texto, imagen, audio. Una imagen vale más que mil palabras, o eso se dice, y debe ser por esto por lo que este tipo de información (las imágenes) son los objetos multimedia más demandados en la actualidad, principalmente en Internet (servicios para compartir fotos personales, redes sociales dedicadas, arte, etc.). Más adelante, en el apartado 2.2, se analizará la tarea de Recuperación Multimedia de Imágenes, y se describirán las técnicas más utilizadas, así como el principal obstáculo que se encuentra al afrontar esta tarea: el problema de la brecha semántica (*multimedia semantic gap*), que se refiere a la complejidad para entender la información que un usuario percibe a partir de las características de bajo nivel de un objeto multimedia.

Se trabaje con documentos textuales o con documentos multimedia, el esquema general de un Sistema de Recuperación de Información completo debe incluir los siguientes componentes:

1. Colección de documentos. Conjunto de documentos que se quiere almacenar, y sobre los que se desea poder realizar consultas para recuperar información.

2. Representación e indexación de la información. Generación de un índice que almacene las características de los documentos de la colección, de modo que permita una posterior búsqueda y recuperación de la manera más eficiente y precisa posible. Para esto, habrá que definir la forma de representar los documentos.

3. Búsqueda según la necesidad de información del usuario (consulta). Interpretación de la consulta con la que el usuario expresa su necesidad de información al sistema, que debe ser capaz de responder con el conjunto de documentos más relevantes a dicha consulta.

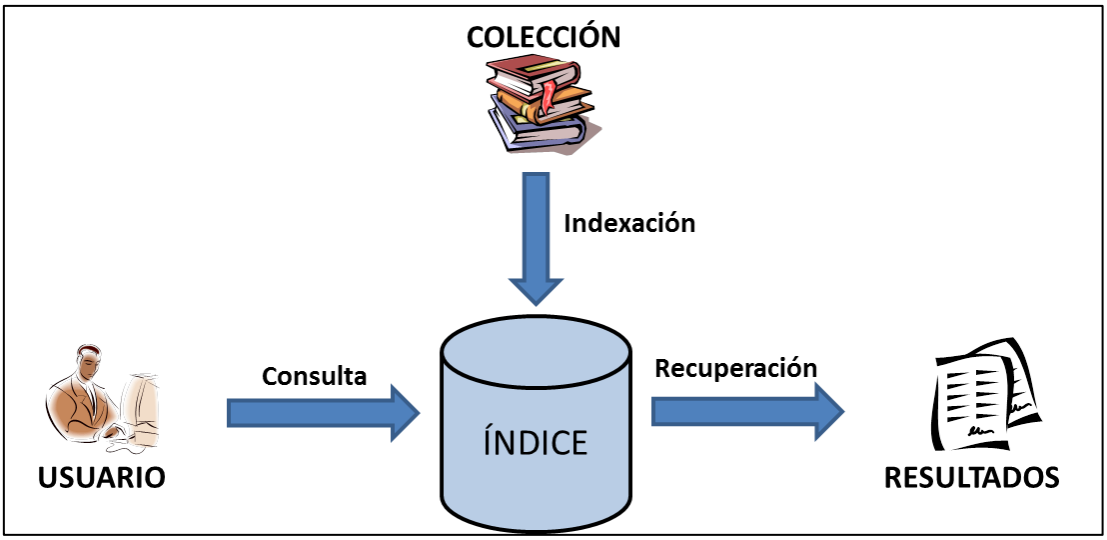

Figura 2.2 Componentes principales de un sistema de Recuperación de Información

Cuando el usuario lance su consulta, el módulo de recuperación se encargará de representar dicha consulta, de la misma manera en la que el módulo de indexación representó los documentos, y de seleccionar aquellos que sean más relevantes para con la consulta introducida (en base a los algoritmos de comparación y modelo de recuperación seguido). Se tratan a continuación algunos aspectos importantes sobre la Recuperación de Información.

2.1.1 Modelos de Recuperación

Los modelos de recuperación de información siguen un esquema consistente en dos pasos:

1) la concepción de un marco de representación para documentos y consultas.

2) la definición de una función de *ranking* que permita cuantificar la similitud entre documentos y consultas. Una función de *ranking* se encarga de asignar puntuaciones (o *scores*) a los documentos en relación a una determinada consulta.

Se muestran a continuación los principales modelos de recuperación utilizados sobre colecciones textuales, a partir de la taxonomía propuesta en (Baeza-Yates and Ribeiro-Neto, 2011).

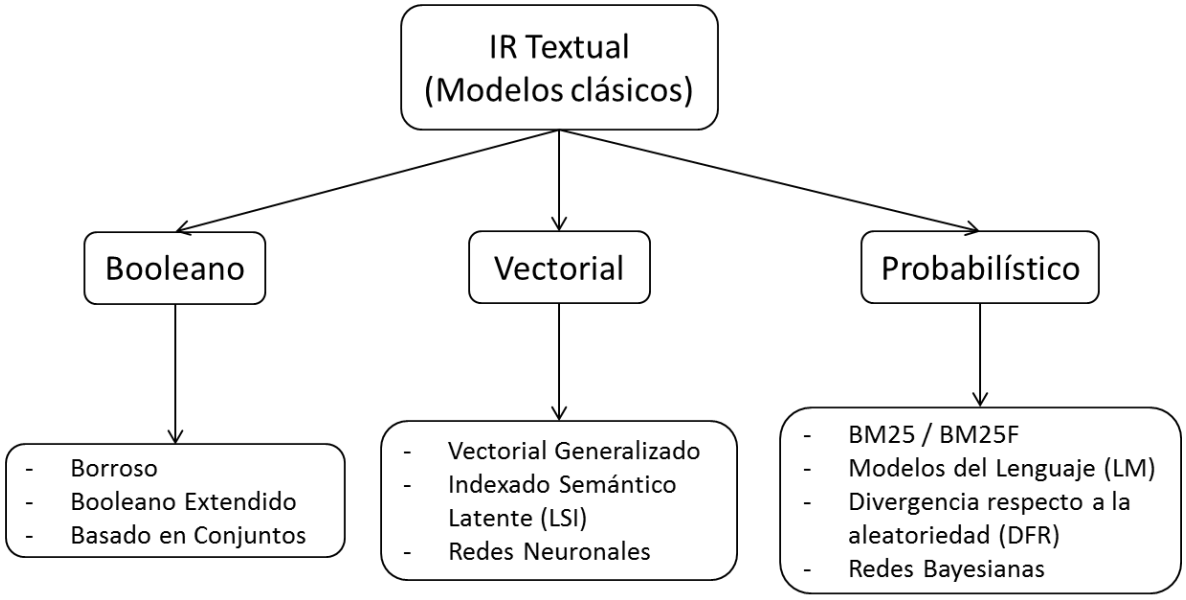

Figura 2.3 Modelos clásicos de Recuperación de Información Textual

A continuación se describen brevemente cada uno de los tres tipos principales de modelos de recuperación, indicando sus principales ventajas e inconvenientes.

2.1.1.1 Modelo Booleano

Se trata del modelo más antiguo (van Rijsbergen, 1999), y está basado en la teoría de conjuntos de George Boole que hace uso del álgebra booleana. Las consultas se definen como expresiones booleanas, sencillas y precisas. Este modelo no es capaz de asignar un valor de relevancia concreto de los documentos con respecto a la consulta, sino que simplemente predice de manera binaria si el documento es relevante o no lo es para la consulta introducida.

Las principales ventajas son su sencillez, y lo rápido e intuitivo que puede llegar a resultar. En cuanto a los inconvenientes, destacar la imposibilidad para generar un conjunto de respuestas ordenadas, debido a la no generación de valores de relevancia/semejanza para los documentos recuperados para cada consulta lanzada.

2.1.1.2 Modelo de Espacio Vectorial (*Vector Space Model*, VMS)

Este modelo fue presentado por primera vez en (Salton, Wong and Yang, 1975). Se trata de un modelo algebraico capaz de representar documentos de texto mediante vectores de términos (palabras) en un espacio multidimensional. Es común utilizar la función coseno entre dos vectores (documentos) para calcular la semejanza entre ellos (aunque en la bibliografía se

pueden encontrar otras medidas de similitud entre vectores). Representando la consulta como un vector más dentro del espacio, puede calculase la relevancia/semejanza (o distancia) de ésta con todos los documentos de la colección. De este modo, se podrá obtener una lista con los documentos relevantes para una determinada consulta, y ordenarlos según su valor de relevancia.

Formalmente, la similitud entre la consulta introducida y los documentos se puede calcular en base a la siguiente fórmula:

$$sim(d_j, q) = \frac{d_j \cdot q}{|d_j| \times |q|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w^2{}_{i,j}} \times \sqrt{\sum_{j=1}^{t} w^2{}_{i,q}}}$$

donde:

- d_j: documento *j*
- q: consulta
- $w_{i,j}$: peso del término i en el documento *j*
- $w_{i,q}$: peso del término i en la consulta q

La principal ventaja de este modelo, aparte de la posibilidad de ordenar los resultados gracias a la función de *ranking*, es que el pesado de los términos mejora la calidad del conjunto respuesta. Por otro lado, la desventaja que presenta es que el modelo asume independencia entre los términos que componen los diferentes documentos.

Las redes neuronales, el modelo del espacio vectorial generalizado (Wong, Ziarko and Wong, 1985) o el indexado semántico latente (LSI, *late semantic indexing*) (Furnas et al., 1988) son subtipos de este modelo de recuperación, muy utilizados.

2.1.1.3 Modelo Probabilístico

Este modelo (Robertson and Jones, 1976), trata de enmarcar el problema de la recuperación de información dentro de un enfoque probabilístico, calculando la probabilidad de que un determinado documento sea relevante para una consulta concreta.

Al igual que en el caso del modelo vectorial, los documentos recuperados quedarán ordenados decrecientemente en función de su probabilidad de ser relevantes. Por otro lado, muestra algunas desventajas, como que el modelo no tiene en cuenta los factores *tf* (frecuencia de cada término), o la ausencia de normalización por la longitud del documento, aspecto que puede influir en la calidad de las respuestas en determinadas colecciones con documentos de longitud variable.

Dentro del modelo probabilístico, pueden destacarse los subtipos mostrados en la figura 2.3: el BM25 o BM25F (por campos), los modelos del lenguaje (LM, *language models*), la divergencia respecto a la aleatoriedad (DFR, *divergence from randomness*), y las redes bayesianas.

BM25 (*Best Match* 25) es una función de *ranking* basada en el modelo probabilístico, utilizada para la asignación de relevancia a los documentos en una determinada búsqueda. Su nombre completo es Okapi BM25 (Robertson et al., 1993), debido al primer sistema que la implementó. Dada una consulta *Q*, con términos $q_1, \dots, q_n$, el valor de relevancia (*score*) asignado mediante la función BM25 para el documentos *D* será:

$$R(q,d) = \sum_{t \in q} \frac{occurs_t^d}{k_1\left((1-b) + b\frac{l_d}{avl_d}\right) + occurs_t^d}$$

donde:

- $occurs_t^d$: frecuencia del término t en el documento d
- l_d: longitud del documento d
- avl_d: longitud media de los documentos de la colección
- k_1, b: parámetros constantes para ajustar la función

Los valores para los parámetros k_1y b suelen depender de las características concretas de cada colección, aunque normalmente se asignan los valores $k_1 = 2.0$ y $b = 0.75$, los cuales se han establecido a partir de los experimentos que durante años se han realizado en las conferencias

TREC[13]. Elegir $b = 0$ sería equivalente a eliminar el proceso de normalización, con lo que la longitud del documento no afectaría al score final. BM25 con $b = 0$ se conoce como BM15, y con $b = 1$ como BM11.

Como extensión de la función de *ranking* BM25 surge BM25F, pensada para documentos estructurados, es decir, formados por campos. En BM25F el peso de un término *t* para un documento *d*, se calcula como una suma ponderada donde a cada campo se le asigna diferente peso (factor de empuje o *boost*) dentro del documento:

$$weight(t,d) = \sum_{c\ in\ d} \frac{occurs_{t,c}^{d} \cdot boost_c}{\left((1-b_c) + b_c \cdot \frac{l_c}{avl_c}\right)}$$

donde:

- l_c: longitud del campo *c*
- avl_c: longitud media para el campo *c*
- b_c: parámetro constante (similar al *b* de BM25)
- $boost_c$: factor de empuje aplicado al campo *c*

La relevancia (*score*) de un documento *d* para una determinada consulta *q* será:

$$R(q,d) = \sum_{t\ in\ q} \frac{weight(t,d)}{\left(k_1 + weight(t,d)\right)} \cdot idf(t)$$

$idf(t)$ se calcula como:

$$idf(t) = log\frac{N - df(t) + 0.5}{df(t) + 0.5}$$

donde:

- N: número de documentos en la colección

[13] http://trec.nist.gov/

- *df*: número de documentos en los que aparece el término *t*

Cuando BM25F se aplica con factores de empuje iguales para todos los campos, la diferencia efectiva entre la aportación de cada campo se deberá a las diferencias encontradas en los valores medios de las longitudes de los campos. En ese caso la importancia de cada campo vendrá marcada por la longitud de los textos contenidos en cada uno de los campos de los documentos.

2.1.1.4 Comentarios

Según (Baeza-Yates and Ribeiro-Neto, 2011), se puede considerar que el modelo booleano es el modelo más débil de los tres. El problema principal que presenta es que no es capaz de reconocer coincidencias parciales entre los documentos de la colección y las consultas, lo que hace que obtenga peores resultados de recuperación.

En cuanto a una comparación sobre el rendimiento de los modelos vectorial y probabilístico, existen varias opiniones. Según autores como Croft, el modelo probabilístico ofrece mejores resultados de recuperación que el vectorial, pero otros como Salton y Buckley afirman que el modelo vectorial es capaz de proporcionar un mejor rendimiento que el probabilístico cuando trabaja con colecciones generales.

Como se verá más adelante, el modelo de recuperación utilizado por el sistema de recuperación de imágenes basado en texto que se utiliza en este trabajo seguirá la aproximación vectorial.

2.1.2 Preprocesamiento de la Información Textual

Independientemente del modelo de recuperación elegido para representar la información y, posteriormente, poder satisfacer las búsquedas de los usuarios, al texto se le suelen aplicar un conjunto de técnicas o preprocesos con el objetivo de hacer más eficiente y preciso el rendimiento global de un sistema de recuperación. Es lo que se llama preprocesamiento de la información textual.

En esta sección se describen brevemente diferentes técnicas de preprocesamiento que suelen aplicarse a los textos antes de su indexación dentro del modelo elegido.

Análisis léxico y caracteres especiales. El análisis léxico de un texto consiste en transformar la cadena de caracteres que forman ese texto en una cadena de palabras, que serán las candidatas para convertirse en términos de indexación. Para lograr este objetivo, la aproximación más común consiste en identificar los espacios en blanco del texto y separar los caracteres en base a ellos.

En esta fase ha de decidirse el tratamiento que se le va dar a determinado tipo de caracteres, como pueden ser: números, acentos, mayúsculas, guiones, signos de puntuación, etc. El objetivo es transformar este tipo de caracteres de forma que faciliten el posterior tratamiento del texto, teniendo en cuenta el modelo de indexación y recuperación que se decida seguir. Por ejemplo, puede decidirse convertir todas las mayúsculas a minúsculas, lo que no debe hacerse si por ejemplo se desea aplicar un proceso de reconocimiento de entidades nombradas, ya que la presencia de esa letra mayúscula será importante para su identificación.

Eliminación de *Stopwords.* Se llama *stopwords* (o palabras vacías) a aquellas palabras que se consideran sin significado, como artículos o preposiciones. En un sistema de recuperación textual este tipo de palabras suelen ser eliminadas para evitar introducir ruido en el proceso de búsqueda y, al mismo tiempo, simplificar el sistema en cuanto a espacio y memoria necesarios para almacenar el índice a crear.

La selección de la lista de stopwords dependerá del sistema de recuperación que se esté desarrollando, y por tanto, no existe una lista única o definitiva. Por ejemplo, en una colección de textos que hablen sobre arte, la palabra “cuadro” puede no tener mucho peso discriminativo entre unos documentos y otros (porque probablemente aparezca varias veces en cada uno) y pueda considerarse como una palabra vacía en el dominio y por tanto ser eliminada para mejorar el rendimiento del sistema. En cambio, esta misma palabra “cuadro” en una colección de información general puede ser muy relevante para diferenciar entre noticias de deportes y de cultura.

También debe tenerse en cuenta cuándo llevar a cabo este proceso de eliminación de stopwords. Por ejemplo, si desea realizarse un análisis sintáctico y extraer los sintagmas nominales presentes en un texto nunca deben eliminarse antes las stopwords, ya que palabras consideradas como vacías semánticamente (“de” o “en”) pueden ser de vital importancia para el proceso sintáctico.

***Stemming* y lematización.** Se llama *stemming* al proceso de reducir una palabra a una raíz (base o *stem*). En cambio, la lematización consiste en convertir una palabra en su forma sin flexionar (lema). Por ejemplo, para la palabra "límite" su raíz o *stem* sería "limit" y su lema "limitar" (puede no coincidir el *stem* y la raíz léxica).

La técnica de *stemming* más utilizada es el conocido algoritmo de Porter, originalmente escrito en 1979 y publicado varias veces con diferentes revisiones (van Rijsbergen, Robertson and Porter, 1980) (Porter, 1980) (Porter, 1997). Resulta muy útil para aquellas lenguas poco flexionadas, por lo que para el caso del español los resultados no siempre son adecuados, aunque suele utilizarse.

Las técnicas de lematización consisten en obtener la forma representante (lema) de todas las posibles flexiones de una palabra: singular para sustantivos, masculino singular para adjetivos, infinitivo para verbos. Para ello se necesita, en general, un diccionario morfosintáctico.

Con el uso de este tipo de técnicas se consigue, por un lado, reducir el tamaño del índice a construir y, por otro, aumentar la cobertura del sistema de recuperación (aunque a la vez suele perderse precisión). El motivo es que muchas palabras distintas serán tratadas por el sistema como si fueran la misma.

2.1.2.1 Expansión de la Consulta

Este proceso consiste en la reformulación de la consulta original con el objetivo de recuperar mejores resultados. Esta técnica puede consistir en añadir sinónimos para los términos de búsqueda y añadirlos a la misma. Por ejemplo, la consulta "fotos de casas" podría expandirse a "fotos de casas hogares", y así podrían recuperarse no solo aquellos documentos que contuviesen el término "casas" sino también en los que apareciese "hogares".

Para expandir las consultas hay técnicas que utilizan lexicones, ontologías, o mediante realimentación por relevancia. La realimentación puede estar basada en información proporcionada por el usuario, pero también por información extraída del conjunto de documentos inicialmente recuperados por el sistema (en este caso se le llama pseudo-realimentación por relevancia).

Con la utilización de esta técnica suelen obtenerse mejores valores de cobertura en la recuperación, aunque en la mayoría de los casos esto repercute en una disminución en cuanto a la precisión del sistema. Según (Carpineto and Romano, 2012) la expansión de la consulta tiene el potencial para superar una de las grandes limitaciones de los sistemas de búsqueda, que es la dificultad que tienen los usuarios para proporcionar descripciones precisas al expresar sus necesidades de información. En la actualidad existen un gran número de técnicas de expansión utilizadas (lingüísticas, basadas en corpus, en *logs*, en recursos externos, etc.), dependiendo del dominio, la colección, o la aplicación concreta de búsqueda.

2.1.2.2 Reconocimiento de Entidades Nombradas

Esta técnica (NER, *Named Entity Recognition*) trata de identificar en los textos entidades nombradas. Una vez identificadas, dichas entidades recibirán un preprocesamiento textual especial. La mayoría de herramientas disponibles para realizar esta tarea (FreeLing, Stilus, Standford, etc.) ofrecen además información acerca del tipo de entidad que es detectado en cada momento: persona, lugar, organización, temporal, etc.

Los sistemas NER pueden estar basados en analizadores de texto (por ejemplo, buscando secuencias de palabras que comiencen con mayúscula), o pueden utilizar grandes bases de conocimiento (como ontologías, lista de lugares geográficos, listas de personalidades, etc.). se trata de un recurso muy utilizado actualmente, que influye positivamente en la calidad de la recuperación.

2.1.2.3 Multilingüismo

El problema del multilingüismo aparece en colecciones que contienen documentos textuales en más de un idioma, o cuando un usuario realiza una consulta al sistema de recuperación de información en un idioma distinto al de los documentos que puede recuperar.

Los sistemas de recuperación de información multilingüe (o CLIR, *Cross-Language Information Retrieval*) son capaces de recuperar documentos independientemente del idioma en el que se haya realizado la consulta y del idioma de los documentos de la colección. El principal problema de estos sistemas será superar la barrera lingüística existente entre el idioma de la consulta y los de los documentos de la colección (Oard, 1997). Los modelos se basan en la traducción de la consulta y/o de los documentos a los idiomas en los que están escritos (Ren and Bracewell, 2009). Uno de los trabajos desarrollados durante la elaboración

de esta tesis está relacionado con el multilingüismo y su aplicación dentro del campo de la recuperación de imágenes basada en texto (Hernández et al., 2011). En dicho trabajo se distinguen tres modelos de indexación:

- tantos índices independientes como idiomas haya, que contendrán la información textual de cada idioma por separado
- un índice expandido que traducirá la información de cada idioma a los demás
- un único índice con la información conjunta de todos los idiomas.

Los resultados obtenidos en el trabajo mencionado muestran cómo el enfoque que mejores resultados obtiene es el que traduce todos los documentos al resto de idiomas, aunque resaltando el inconveniente del elevado coste computacional requerido para completar la tarea.

2.2 Recuperación Multimedia de Imágenes

Esta tesis centrará su investigación en el campo de la recuperación multimedia de imágenes, haciendo uso de los modos visual (las imágenes propiamente dichas) y textual (texto asociado a las imágenes).

La gran cantidad de imágenes disponibles hoy en día tanto en la Web, como en colecciones personales o privadas, ha hecho que la tarea de recuperación de imágenes haya alcanzado un interés realmente alto dentro de la comunidad científica. Hechos como la aparición de la cámara digital o los dispositivos móviles, sobre todo los teléfonos inteligentes (o *smart phones*), junto con el creciente desarrollo de Internet en aplicaciones de compartición de fotos (Flickr, Instagram, Picassa, etc.) y en redes sociales (Twitter, Facebook, Pinterest, etc.), han elevado el número de imágenes que de algún modo hay que gestionar. Todos estos factores han abierto un sinfín de posibilidades para los diseñadores de sistemas de búsqueda de imágenes en el mundo real (Datta et al., 2008).

Algunos datos numéricos con los que hacerse una idea de la magnitud de la cantidad de imágenes con las que se enfrentan los sistemas de almacenamiento y búsqueda pueden ayudar a entender la importancia del problema. Por ejemplo en Facebook se suben cada día 300.000 millones de fotografías (según http://newsroom.fb.com/Products/Photos-and-Video-15b.aspx,

30/04/2012). En Flickr (según *2010 Internet Statics*, http://fritzboyle.com/2010-internet-statistics/) están alojadas 5.000 millones de imágenes, y se suben cada mes 130 millones más. Más datos: en Facebook se suben cada día 250 millones de fotos (según http://thesocialskinny.com/100-more-social-media-statistics-for-2012/). Según información de http://royal.pingdom.com/2012/01/17/internet-2011-in-numbers/, el número medio de fotos subidas por segundo a Instagram es de 60 (http://blog.instagram.com/tagged/yearinreview), donde se alojan unos 400 millones de ellas. Se estima que a mediados del año 2011 Facebook tiene unos 100.000 millones de fotos (http://www.onlinemarketing-trends.com/2011/03/facebook-photo-statistics-and-insights.html). Se suben a Flickr 4.5 millones cada día (http://advertising.yahoo.com/article/flickr.html). A fecha de Agosto de 2011 hay 6.000 millones de fotos hospedadas en Flickr (http://blog.flickr.net/en/2011/08/04/6000000000/).

La recuperación de imágenes ofrece nuevos retos y tiene unas características bastante distintas si se compara con la recuperación de documentos textuales. Estas diferencias pueden observarse en la forma en que se expresa una consulta, los métodos utilizados para la recuperación (Rui, Huang and Chang, 1999), los tipos de consultas, cómo se decide la relevancia, el papel del usuario durante el proceso de búsqueda, y en general, cualquier diferencia cognitiva entre la interpretación de los modos visual y textual.

La búsqueda de imágenes más utilizada hasta hace poco tiempo es la basada en las descripciones textuales de las imágenes, siguiendo las técnicas de recuperación de información. Pero en el mundo real no siempre se dispone de información textual de calidad asociada a las imágenes. Cuando no se dispone de este tipo de anotaciones, la única alternativa es realizar las búsquedas en base al contenido de las imágenes, esto es, a su información visual. La forma de representar las características visuales de una imagen es por medio de los descriptores de bajo nivel, relacionados con aspectos de las imágenes como su color, su textura o su forma.

El "gap semántico" en multimedia, *multimedia semantic gap*, o brecha semántica (Smeulders et al., 2000), se refiere a la complejidad para entender la información que un usuario percibe a partir de las características de bajo nivel de un objeto multimedia (imágenes, vídeos, audio). Las imágenes (y los objetos multimedia en general) se representan mediante una serie de características de bajo nivel, mientras que los usuarios buscan dichas imágenes expresándose

mediante conceptos semánticos de alto nivel. En general, puede entenderse el *semantic gap* como la diferencia entre las descripciones de un mismo objeto mediante diferentes representaciones.

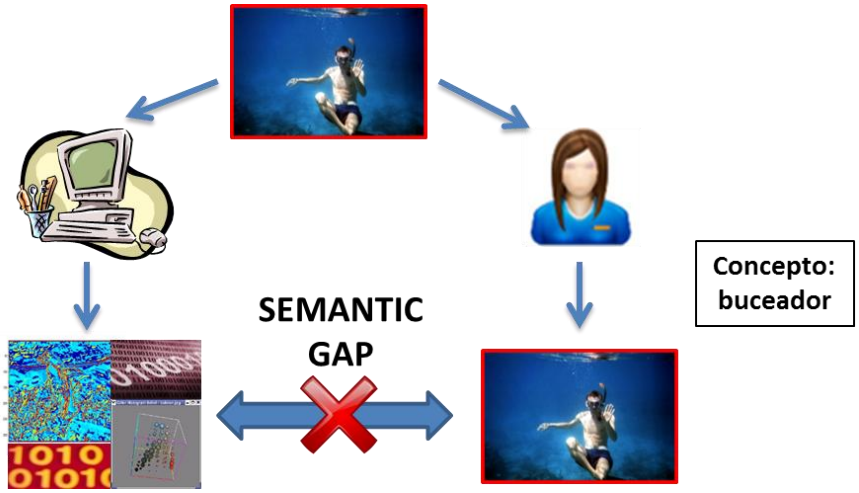

Figura 2.4. *Multimedia Semantic Gap*

La fusión multimedia intenta utilizar las diferentes medias como información complementaria para aumentar la precisión de los resultados recuperados. En concreto, en el caso de la recuperación multimedia de imágenes, el gap semántico es la falta de correspondencia entre la información presente en las características visuales (histogramas de color, forma, textura, etc.) y la interpretación de estos datos por parte del usuario.

La siguiente imagen, inspirada en (Baeza-Yates and Ribeiro-Neto, 2011), ayuda a entender el problema del *semantic gap* como la brecha entre una señal multimedia (contenido) y su significado (semántica).

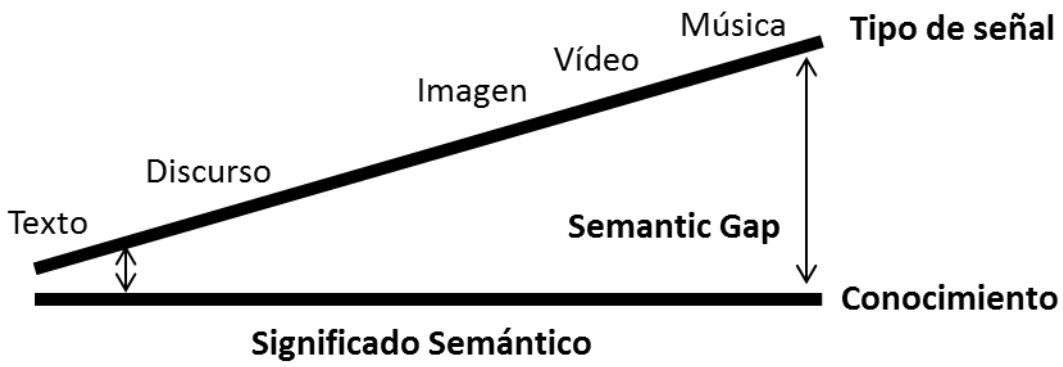

Figura 2.5 Escala semántica en Multimedia

La imagen muestra una hipotética escala indicando la carga semántica de los diferentes tipos de información multimedia. Se observa cómo la diferencia entre el significado real o semántico de un objeto multimedia crece según el tipo de señal. La mayor carga semántica se encuentra en el texto, mientras que las modalidades de imagen, vídeo y audio (música o habla/discurso) llevan implícita un nivel menor de semántica, con lo que la brecha (*semantic gap*) se hace cada vez más grande y, por lo tanto, más difícil de superar. Se incluyen ejemplos que aclaran este concepto en la segunda parte de esta memoria de tesis.

Para el caso concreto de la tarea de recuperación multimedia de imágenes, se propondrá en esta tesis el uso de una aproximación de fusión multimedia para abordar el problema del *semantic gap*, y se mejorarán los resultados de los enfoques monomodales (TBIR y CBIR) sacando provecho de la complementariedad existente entre la información proveniente de ambos modos (texto e imagen).

2.2.1 Recuperación de imágenes basada en texto (TBIR)

Este tipo de recuperación hace uso de la información textual asociada a las imágenes y, mediante alguno de los modelos de recuperación textual descritos en el apartado 2.1.1, es capaz de llevar a cabo la recuperación de un conjunto de imágenes que satisfagan las necesidades de información del usuario (que también deberán expresarse en modo textual).

Aunque este enfoque ha sido el utilizado por gran cantidad de herramientas comerciales y motores de búsqueda (incluido Google), presenta un problema evidente: las imágenes han de estar anotadas textualmente, y de forma “adecuada” a la expresión que un usuario use en la consulta para referirse a ella. La ingente cantidad de imágenes existentes hoy en día hace que

la anotación textual sea una misión prácticamente imposible. Aún en el caso de que la información textual asociada a las imágenes sea generada a partir de su contexto (pies de foto, leyendas, título de sección/página, párrafo más cercano, etc.), la búsqueda también puede resultar imprecisa ya que no siempre esta información contextual es acertada (Jaramillo and Branch Bedoya, 2008).

2.2.2 Recuperación de imágenes basada en contenido (CBIR)

Esta aproximación surge en la década de los 90 debido a la imposibilidad de disponer de anotaciones textuales para todas las imágenes, y de su imprecisión y subjetividad en caso de tenerlas, cuando las colecciones alcanzan tamaños de miles de imágenes (Rui, Huang and Chang, 1999). La idea es no necesitar esas anotaciones textuales, sino representar, indexar y recuperar las imágenes en base a su propio contenido (utilizando características como su color o su textura). Los sistemas CBIR (*Content-Based Image Retrieval*) también son conocidos como CBVIR (*Content-Based Visual Information Retrieval*) o QBIC (*Query By Image Content*).

La extracción de características visuales de bajo nivel es la base de la recuperación de imágenes basada en contenido. Una característica se define para capturar una determinada propiedad de una imagen. Se describen a continuación las características visuales principales que pueden ser extraídas de una determinada imagen (Rui, Huang and Chang, 1999) (Baeza-Yates and Ribeiro-Neto, 2011) (Datta et al., 2008).

Color. Es una característica visual relativamente robusta, independiente del tamaño y resolución de la imagen, o de su orientación (ángulo de visión). La semejanza entre dos imágenes dadas se calculará en base a la comparación realizada entre los histogramas de color de las imágenes de la colección y el de la imagen de consulta. Un histograma de color c_i para una imagen I se define como la probabilidad de que el pixel *p* seleccionado aleatoriamente de la imagen I tenga el color c_i, esto es:

$$h_I(c_i) = P(color(p) = c_i | p \in I)$$

Textura. La textura de una imagen se corresponde con un fenómeno perceptivo, fácilmente detectable por los humanos, pero difícil de describir de manera matemática. Por ejemplo, la hierba y los pétalos de rosa se diferencian en cuanto a textura por su suavidad, así como por

sus patrones. La textura se refiere a los patrones visuales que tienen propiedades de homogeneidad que no son resultado de la presencia de un único color o intensidad (Smith and Chang, 1996). Se trata de una medida de los elementos repetitivos en la imagen. Caracteriza los patrones que se repiten sobre la intensidad de la imagen que son demasiado finos para ser diferenciados como objetos independientes.

Su papel en la recuperación de imágenes, sobre todo en el caso de estar trabajando en un dominio específico (por ejemplo, el médico), es importantísimo, debido a su estrecha relación con la semántica subyacente en dichos casos.

Forma. Se trata de un atributo clave de las regiones de imágenes segmentadas, y su eficiencia y robustez juegan un importante papel en la recuperación. Dependiendo de la aplicación en la que se esté trabajando, puede necesitarse que la representación de la forma sea invariante a la translación, o a la rotación, o a la escala.

Las representaciones para la forma pueden dividirse en dos categorías: las basadas en límites y las basadas en región. Las primeras utilizan el límite exterior de la forma, mientras que las segundas se basan en la región completa de la forma. Las dos representaciones más exitosas para estas dos categorías son el descriptor de Fourier (utilizando el límite de la transformada de Fourier) y los momentos invariantes (haciendo uso de los momentos basados en la región, que son invariantes a transformaciones).

Distribución del color. Combina características de color con relaciones espaciales para mejorar los resultados obtenidos por la característica de color global, que aunque proporciona un poder de discriminación razonable para la recuperación de imágenes, tiende a dar demasiados falsos positivos cuando la colección es grande.

De igual modo que con el color, también se puede extraer la distribución de otras características visuales (como la textura) para facilitar una recuperación de imágenes más avanzada.

Puntos principales. Se trata de una técnica que busca características de la imagen que sean constantes sobre diferentes escalas. Son especialmente robustos a cambios de iluminación, de posición de cámara, y de ángulo del objeto. Analiza la apariencia de la imagen en los puntos que son especialmente distintivos (como las esquinas).

2.2.2.1 Medidas de similitud en CBIR

Para calcular la semejanza entre dos imágenes en base a las características de bajo nivel mostradas en el apartado anterior, el método más utilizado usa una medida de distancia entre las imágenes. Una distancia igual a cero implicará una coincidencia exacta, esto es, significará que las imágenes son iguales. El valor de distancia obtenido gracias a esta medida entre el conjunto de imágenes donde se está buscando y la imagen de ejemplo de la consulta permitirá ordenar las imágenes recuperadas por un sistema CBIR según su semejanza.

Se muestran a continuación varias medidas o técnicas utilizadas comúnmente para calcular la similitud entre dos imágenes (Kekre, Mishra and Kariwala, 2011) utilizando un algoritmo basado en contenido o en características visuales:

Distancia Euclídea. En este caso la métrica utilizada para calcular la distancia entre los vectores de características que describen dos imágenes es la métrica Euclídea. Cuando se obtienen valores altos según esta medida se está indicando que la distancia entre las caracterizaciones de las imágenes es alta y, por lo tanto, que la semejanza entre ellas es baja. La fórmula para calcular la distancia Euclídea en un espacio n-dimensional sería:

$$d_E(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Error cuadrático medio. El error cuadrático medio (MSE, *Mean Square Error*) es otra técnica utilizada para medir distancias. Valores bajos de MSE indicarán similitud entre las imágenes. El MSE en estadística es una manera de cuantificar la diferencia entre el valor implícito de un estimador y su valor verdadero. Si se lleva esta definición al campo de aplicación del reconocimiento de imágenes, los estimadores serían las imágenes de la colección y el valor a ser estimado sería la imagen ejemplo de la consulta con la que se está comparando. Si se toma X como parámetro del estimador con respecto a X' como el valor estimado, el error cuadrático medio será el dado por la siguiente ecuación, que calcula la esperanza (E) del cuadrado de la diferencia entre el estimador y el valor estimado (se usa el cuadrado para tener en cuenta de igual manera los errores en ambas direcciones):

$$MSE(X') = E(X' - X)^2$$

Suma de diferencias absolutas. Es un algoritmo para encontrar la correlación entre dos imágenes. Toma las diferencias absolutas entre cada pixel de la imagen original con respecto al pixel correspondiente de la imagen utilizada para la comparación. Al igual que la distancia Euclídea y que el MSE, valores bajos de esta suma de diferencias absolutas significará semejanza entre las imágenes comparadas. La precisión de este método se puede ver afectado por factores como la iluminación, la forma o el tamaño.

Mahalanobis. Fue introducida en (Mahalanobis, 1936), para determinar la similitud entre dos variables aleatorias multidimensionales teniendo en cuenta la correlación entre las variables aleatorias (a diferencia de la distancia Euclídea).

Formalmente, esta distancia entre dos variables aleatorias (x, y) con la misma distribución de probabilidad y con matriz de covarianza Σ, se define como:

$$d_M(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$$

2.2.3 Recursos externos auxiliares

Existen diversos recursos externos que han sido tenidos en cuenta durante el desarrollo de este trabajo como apoyo para la tarea de recuperación multimedia de imágenes, como es el caso de los que se describen a continuación.

Wikipedia. Este recurso fue utilizado durante la experimentación con la colección de imágenes de Wikipedia proporcionada por los organizadores de la tarea de recuperación en las ediciones de ImageCLEF 2010 y 2011. Wikipedia, según su propia autodefinición, es una enciclopedia libre y políglota con más de 20 millones de artículos en 282 idiomas y dialectos (17 de ellos con más de 300.000 artículos), redactados por voluntarios de todo el mundo. Fue iniciada en 2001 y es actualmente una popular obra de consulta en Internet.

Como se describe en el apartado 4.2.1.3, la colección mencionada está formada por imágenes extraídas de artículos de Wikipedia. Un grupo de experimentos (mostrados en el apartado7.1.2) evalúa la conveniencia de hacer uso de la información textual proporcionada en los artículos a los que pertenecen las imágenes de la colección para enriquecer las anotaciones textuales o metadatos asociados a cada imagen. Como se detallará en el capítulo de experimentación mencionado, el enriquecimiento de la información textual asociada a las

imágenes haciendo uso de los títulos y categorías (campos <*title*> y <*category*>) de los artículos donde se encuentran las imágenes, redundará en una recuperación basada en texto (TBIR) mucho más precisa y con unos resultados mejores.

El baseline textual desarrollado en esta tesis utiliza este recurso para enriquecer la información asociada a las imágenes.

ImageNet. Se trata de una ontología de imágenes organizada de acuerdo a la jerarquía de WordNet[14] (actualmente solo de los nombres), en la que cada nodo de la jerarquía está descrito por una media de 500 imágenes. La idea de este recurso es poder resultar de utilidad a la comunidad investigadora en el campo CBIR, ya que las imágenes no están anotadas textualmente, pero sí clasificadas en los *synsets* de WordNet.

En la Figura 2.1 se muestra una captura de una rama de la estructura de ImageNet que va desde la raíz a las hojas para el subárbol correspondiente al término *mamífero* (*mammal*). Para cada *synset* de WordNet se muestra un conjunto de 9 imágenes seleccionadas aleatoriamente:

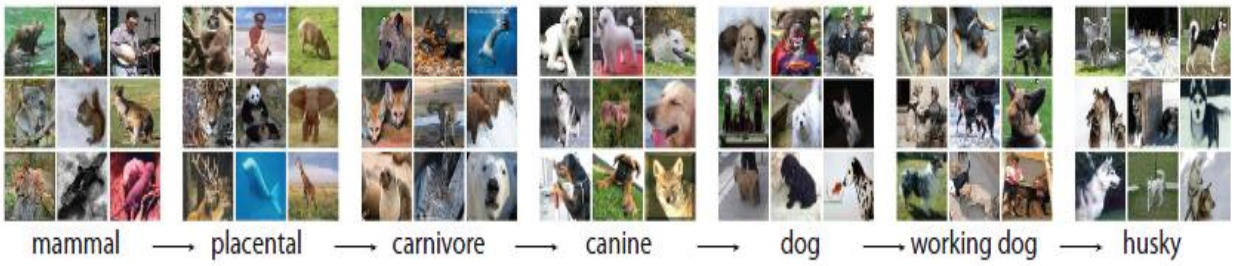

Figura 2.6. Ejemplo de imágenes en ImageNet (para el synset "mamífero")

Cabe destacar tres posibles aplicaciones en las que el uso de ImageNet podría resultar beneficioso (Deng et al., 2009): reconocimiento de objetos en imágenes, clasificación de imágenes, y *clustering* automático de objetos.

En cuanto a la relación entre las categorías visuales y semánticas implícita en la ontología, en (Deselaers and Ferrari, 2011) se observa que la similitud visual crece con la similitud semántica, y que las categorías visuales son separables a lo largo de límites semánticamente

[14] http://wordnet.princeton.edu/

definidos. Para los objetivos y experimentos de esta tesis no se ha encontrado utilidad, ya que las imágenes no están anotadas textualmente.

Delicious. Es un servicio de gestión de marcadores sociales en la Web. Su utilización permite guardar los marcadores personales en el sistema y organizarlos en categorías con un sistema de etiquetado denominado *folksonomías*. El sistema permite compartir estos marcadores con otros usuarios.

En uno de los trabajos realizados durante el desarrollo de esta tesis (Cigarrán Recuero et al., 2011) se utilizó el corpus *DeliciousT140 Dataset* (Zubiaga, 2011) para expandir las consultas en la tarea de etiquetado de vídeos propuesta en la edición de 2011 de MediaEval, consistente en clasificar un conjunto de vídeos anotados en base a una lista de etiquetas predefinidas. El corpus mencionado, generado a partir de Delicious, se compone de 144.574 URL únicas, todas ellas con sus correspondientes etiquetas sociales recuperadas desde Delicious. Para aplicar la expansión de consulta mencionada, se analiza la información en el corpus *DeliciousT140* para encontrar las etiquetas sociales que coocurren con cada una de las etiquetas de la lista de clasificación. Estas coocurrencias se ordenan en función de la frecuencia de cada una, y a partir de esas listas se proponen los términos de expansión. Este recurso no se ha utilizado en los trabajos incluidos en esta memoria, ya que no se han seguido aproximaciones de recuperación textual de imágenes basadas en expansión de la consulta para evitar la inclusión de posible ruido en la lista de resultados.

Large-Scale Concept Ontology for Multimedia (LSCOM). Ontología desarrollada dentro del proyecto Mediamill, y formada por un conjunto de 449 conceptos semánticos (Naphade et al., 2006), que incluyen diversas categorías como objetos, actividades/eventos, localizaciones o personas. En la siguiente figura se muestra una versión reducida de LSCOM:

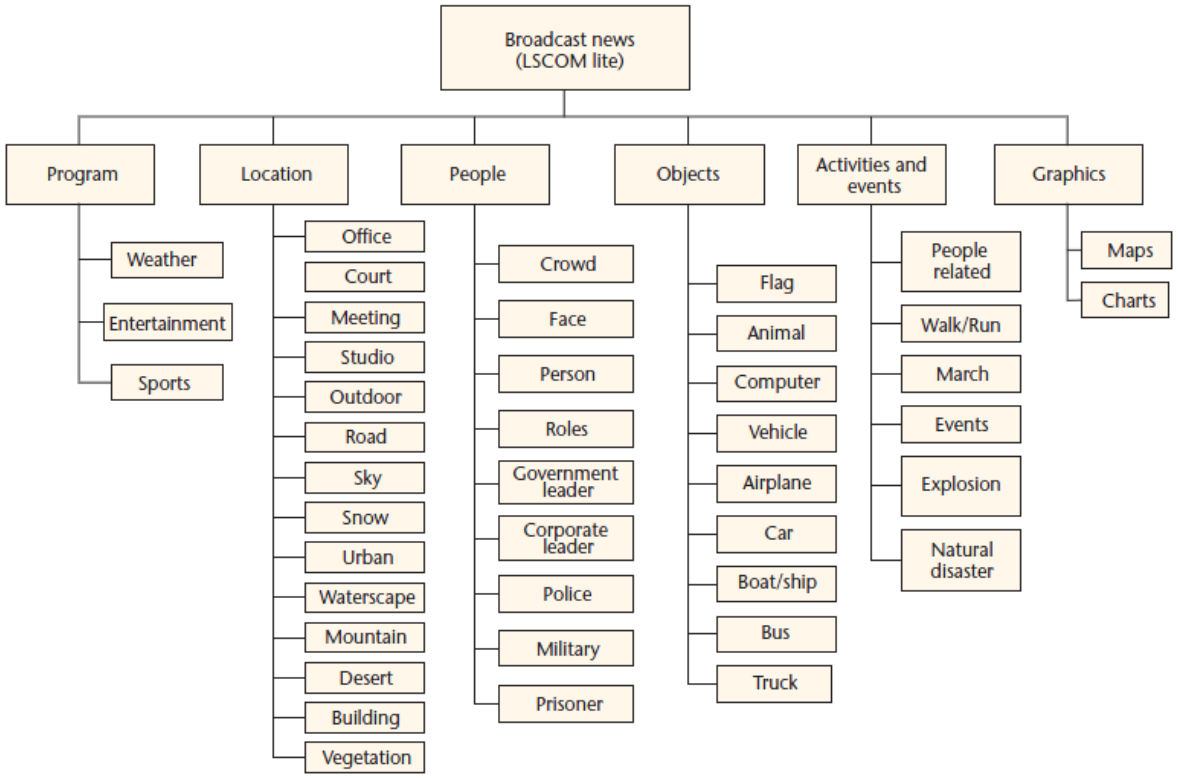

Figura 2.7. Versión reducida de la taxonomía LSCOM

En (Snoek et al., 2006) se afronta el reto de la detección automática de 101 conceptos semánticos dentro del ámbito multimedia, concretamente trabajando con indexado de vídeos, con una ontología de menosr conceptos, pero más utilizada en la bibliografía.

En (Snoek and Smeulders, 2010) también se hace uso de esta ontología para trabajar en la tarea de la búsqueda de imágenes basada en la detección de conceptos visuales, tratando así de resolver el problema de la brecha semántica (*semantic gap*). Para ello siguen un esquema basado en tres pasos: 1) extraer las características visuales de bajo nivel de las imágenes (color, forma, textura, etc.), 2) proyectar los descriptores obtenidos sobre un conjunto de palabras, y 3) utilizar algoritmos de aprendizaje automático (*machine learning*) para convertir las palabras visuales en alguno de los conceptos semánticos de la ontología, asignando una probabilidad a cada uno de los conceptos existentes. Estas probabilidades se utilizarán para ordenar las imágenes en función de la presencia de los conceptos.

La ontología presentada también fue utilizada en uno de los trabajos realizados durante el desarrollo de esta tesis para anotación y expansión de la consulta dentro de una tarea de recuperación multimedia de imágenes (Granados et al., 2012). Un subconjunto de conceptos

de la ontología potencialmente presentes en las imágenes de la colección de evaluación fueron utilizados para crear un conjunto de términos relacionados que fueron empleados para anotar los metadatos asociados a las imágenes, y para expandir las consultas.

En los experimentos de recuperación textual de imágenes desarrollados en esta tesis no se hace uso de esta estrategia, tratando de hacer lo más automática posible la fase de recuperación.

Capítulo 3 Fusión Multimedia

Se dedica este capítulo a definir y describir detalladamente la fusión multimedia, indicando los aspectos principales a tener en cuenta cuando se aborda una tarea de este tipo. Se definen los principales niveles de fusión, y se incluye una clasificación de los métodos de fusión multimedia presentes en distintas aplicaciones, así como la descripción de diferentes algoritmos de fusión tardía, diferenciando entre aquellos basados en la agregación de *scores* o valores de relevancia, y los basados en sistemas de votación. También se realiza una profunda revisión de las técnicas de fusión multimedia aplicadas dentro del escenario concreto de esta tesis, la tarea de recuperación multimedia de imágenes. Finalmente, se trata el proceso de normalización de listas de resultados, describiendo las principales técnicas existentes.

3.1 Introducción

Cualquier concepto semántico se describe mejor si para ello se utilizan diferentes fuentes de información (Muller, Clough and Desealaers, 2010). Por ejemplo, en el campo de la medicina un diagnóstico tendrá mayor confianza si los resultados de laboratorio, la historia clínica del paciente, los posibles exámenes radiológicos, etc., son tenidos en cuenta conjuntamente y convergen hacia una misma conclusión. De esta idea surge la Fusión Multimedia, con el objetivo de aprovechar la información disponible desde los distintos modos existentes en un objeto multimedia.

Para la tarea en la que se centra el trabajo de esta tesis, la recuperación multimedia de imágenes, la fusión tratará de aprovechar tanto la información visual de la imagen, como la textual disponible en las anotaciones o metadatos asociados, combinándolas de la mejor

manera posible para explotar las ventajas de cada una y la complementariedad existente entre ellas. El principal reto de la fusión multimedia es ayudar a resolver el problema del *semantic gap*, tratando que en el proceso de recuperación intervengan tanto los aspectos textuales como los visuales para hacer lo más efectiva posible la colaboración entre ambos modos. Se trata de mejorar, en base a la fusión, el rendimiento de la búsqueda obtenido por las aproximaciones monomodales de manera independiente (TBIR y CBIR). Aunque es conocido que las técnicas puramente visuales obtienen resultados bastante más bajos que los enfoques textuales, los métodos basados en el contenido de las imágenes (modo visual) pueden potencialmente ayudar a mejorar la precisión de los resultados aportando información adicional a la textual (Lew et al., 2006).

En general, cuando se utiliza un enfoque de fusión multimedia (Atrey, Hossain and Kankanhalli, 2010) se han de tener en cuenta varios aspectos:

1) la asincronía en la información disponible proveniente de distintas fuentes, así como su diferente formato,
2) la no independencia entre las modalidades (correlación). Las modalidades independientes pueden proporcionar información adicional en el momento de tomar una decisión,
3) la confianza en las distintas fuentes o modos, que puede variar en función de la tarea a realizar,
4) las restricciones de coste y disponibilidad de proceso relacionado con cada modo y su función.

En cualquier escenario en el que se aborde una tarea de análisis multimedia habrá que intentar alcanzar un equilibrio entre los aspectos anteriores.

3.2 Niveles de Fusión

En la bibliografía (Atrey, Hossain and Kankanhalli, 2010), se diferencian principalmente dos estrategias o niveles de fusión: a nivel de características (*early fusion*, o fusión temprana) y a nivel de decisiones (*late fusion*, o fusión tardía). Cuando se sigue una estrategia que combina estas dos aproximaciones, se habla de fusión híbrida. Se describen a continuación estos niveles de fusión multimedia, indicando las ventajas y desventajas de cada uno de ellos y algunos ejemplos de aplicación. Se describirá también el nivel de *transmedia fusion*,

diferenciado de los anteriores en (Clinchant, Csurka and Ah-Pine, 2011), dentro del que se enmarcaría el modelo de fusión multimedia presentado en este trabajo.

3.2.1 Fusión a nivel de características

También llamada fusión temprana o *early fusion*, consiste en combinar las características extraídas de cada fuente de información para posteriormente realizar la fase de análisis en función del conjunto global de características combinadas.

Las características son las utilizadas para representar la información obtenida desde cada modo disponible, esto es, los descriptores de cada modalidad de información. Por ejemplo, en el campo de la recuperación de imágenes anotadas, se tienen por un lado características textuales (términos, frecuencias, etc.), y por otro las características visuales de bajo nivel (descriptores de color, de forma, etc.). Para otro tipo de tarea de análisis multimedia como el reconocimiento de caras, podrían considerarse características como el color de la piel por un lado, y de movimiento por otro. Toda esta información extraída de los distintos modos se puede considerar conjuntamente, como un único vector global de características, que será la entrada para el sistema o módulo de detección de caras (o de recuperación de imágenes en el primer ejemplo). Esto puede verse más claramente en la Figura 3.1, donde F_n hace referencia al conjunto de características extraídas desde el modo *n*, la "Fusión de características" combina la información procedente desde cada modo, y el "Módulo de decisión" toma la decisión final D:

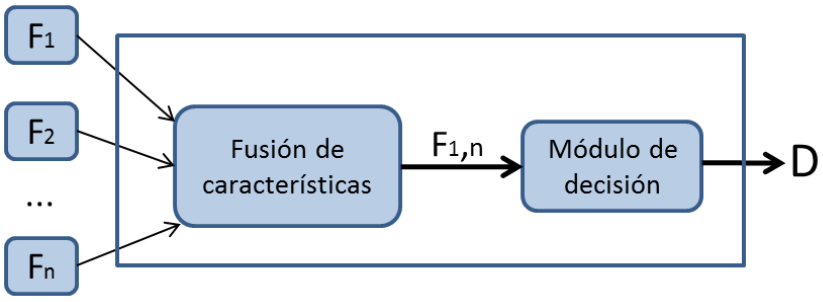

Figura 3.1. Fusión a nivel de características

La principal ventaja de este nivel de fusión es que en él se puede aprovechar la correlación entre las características de los diferentes modos implicados en la tarea. Otro punto positivo importante es que con esta estrategia solo es necesaria una fase de análisis o aprendizaje, en

función del vector global de características combinadas (Snoek, Worring and Smeulders, 2005).

Entre los inconvenientes de este nivel (en aplicaciones en tiempo real) hay que destacar la dificultad para gestionar la sincronización temporal entre características, ya que cada una puede tener distintos tiempos de extracción. Por ejemplo, en la tarea de recuperación de imágenes, la información textual se obtiene casi de manera inmediata tras leer los ficheros de metadatos (por ejemplo), mientras que los descriptores visuales de color necesitan construir sus histogramas (por ejemplo), lo que es más costoso en tiempo. Otra desventaja de esta estrategia es que hay que definir un formato común de representación para las características de los distintos modos, lo cual es una tarea complicada por la heterogeneidad de los descriptores de cada modalidad.

Varios ejemplos de aplicación de este nivel de fusión para la tarea de recuperación de imágenes en el foro *ImageCLEF* se describen en el apartado 3.5 de esta memoria.

3.2.2 Fusión a nivel de decisión

En esta aproximación, también conocida como fusión tardía o *late fusion*, se generan decisiones o resultados de manera independiente a partir de las características individuales de cada modalidad, obteniéndose un resultado desde cada modo. Posteriormente, estas decisiones locales se combinan mediante un módulo de fusión obteniendo la decisión o resultado final. En el escenario de la recuperación de imágenes, cada modo (el textual y el visual) generará una lista independiente de imágenes relevantes para la consulta en función de sus propias características o descriptores. A continuación estas listas de resultados (decisiones de cada modo) son combinadas mediante un módulo de fusión que generará la lista fusionada final de imágenes resultado.

Se observa gráficamente este proceso en la Figura **3.2**, donde hará falta un "Módulo de decisión" para las características de cada modo (F_n) que generarán una decisión individual cada una (D_n), las cuales serán combinadas por el módulo de "Fusión de decisiones" antes de proporcionar la decisión final (D):

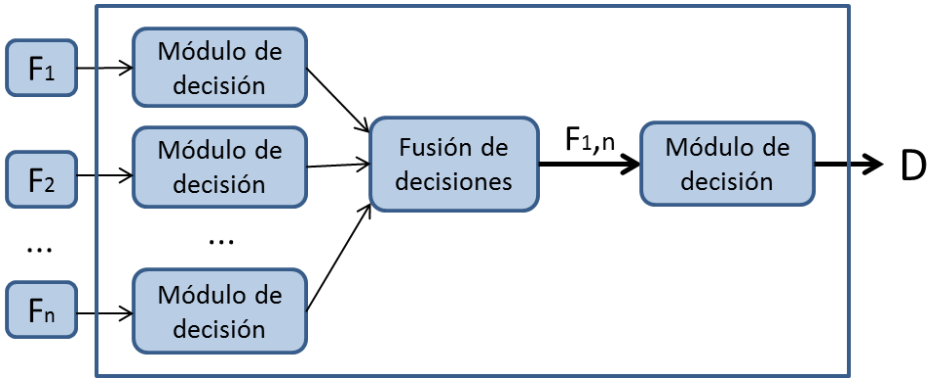

Figura 3.2 Fusión a nivel de decisión

La principal ventaja de esta estrategia es que no tiene que enfrentarse con las distintas representaciones de las características de cada modo, ya que la combinación se realiza a nivel de decisiones y éstas pueden representarse con el mismo formato sin dificultad. Otra ventaja es la escalabilidad en cuanto al número de modalidades que pueden tenerse en cuenta en el proceso de fusión multimedia. Además, para cada modalidad podrán utilizarse los métodos más adecuados para cada caso (mayor flexibilidad), como por ejemplo modelos ocultos de Markov para el audio y máquinas de vectores soporte (SVM) para las imágenes. Además permite que la combinación se vea afectada explícitamente por la confianza o el tipo de información reflejada en la consulta (en un entorno ruidoso, el audio tendrá menor influencia que la transcripción, por ejemplo).

En cuanto a las desventajas, no es posible utilizar directamente la información sobre la correlación entre modalidades como en el caso de la fusión temprana. Por otra parte, el hecho de necesitar tras una fase de decisión previa para cada modalidad, otra de decisión, puede hacer que el proceso sea más costoso en cuanto a tiempo y recursos según el número de medios, aunque más simple a nivel de presentación (fusión de listas de resultados).

Un ejemplo de aplicación de esta aproximación puede verse en (Iyengar, Nock and Neti, 2003), donde se lleva a cabo la fusión de las decisiones obtenidas desde un sistema de detección de caras y otro de reconocimiento del habla. Un caso de fusión tardía en la tarea de recuperación multimedia de imágenes es (Escalante et al., 2008), donde se combinan los resultados de texto e imagen mediante una función de agregación que calcula la media. Otras aproximaciones pueden verse en (Caicedo et al., 2010).

3.2.3 Fusión transmedia

Este nivel de fusión, diferenciado de los anteriores en el trabajo de (Clinchant, Csurka and Ah-Pine, 2011), se basa en procesos de propagación que actúan como mecanismos de pseudo-relevancia entre modos, en vez de aplicar funciones de agregación como en la mayoría de los métodos de fusión tardía. La idea principal de estos procesos es utilizar una de las modalidades para recolectar imágenes relevantes (por ejemplo las K más cercanas o parecidas a la imagen o imágenes de consulta), para posteriormente utilizar la otra modalidad y agregar su información o reordenar por relevancia. Ejemplos de este tipo de aproximación son (Jeon, Lavrenko and Manmatha, 2003), (Maillot, Chevallet and Lim, 2006) y (Ah-Pine et al., 2009).

En (Maillot, Chevallet and Lim, 2006), dentro de una tarea de recuperación multimedia de imágenes, se utilizan técnicas de realimentación por pseudo-relevancia (*pseudo-relevance feedback*) para expandir las consultas (*query expansion*) en ambas direcciones: la consulta textual se expande con el texto asociado a las primeras imágenes recuperadas por el módulo visual, y la consulta visual se expande con las primeras imágenes recuperadas por el módulo textual. Es lo que se denomina realimentación por pseudo-relevancia entre modalidades (*intermedia pseudo-relevance feedback*). Adicionalmente, añaden una fase final de reordenamiento (*re-ranking*) basado en la apariencia visual.

En (Jeon, Lavrenko and Manmatha, 2003) se propone una aproximación automática para anotar y recuperar imágenes en base a un conjunto de entrenamiento. La idea que persiguen es, asumiendo que las regiones de una imagen se pueden describir utilizando un pequeño vocabulario obtenido aplicando *clustering* a las características de la imagen, generar una palabra mediante modelos probabilísticos a partir del conjunto de entrenamiento. Estas palabras serán utilizadas para anotar automáticamente y posteriormente recuperar las imágenes.

Otra forma de fusión entre dos modalidades de información (por ejemplo, texto e imagen) es el llamado método *image re-ranking*, que se divide en dos fases: primero se realiza una búsqueda textual que devuelve una lista de imágenes recuperadas (según su relevancia o similitud textual) y, a continuación, las imágenes recuperadas textualmente se reordenan en función de la similitud visual. Dicho de otra manera, el método *image re-ranking* restringe al

sistema visual a buscar solo entre el conjunto de imágenes recuperadas por una búsqueda textual inicial. Esta aproximación puede formularse del siguiente modo:

$$S_{rerank}(q,d) = I_{\{d \in KNN_t(q)\}} s_v(q,d)$$

donde $KNN_t(q)$ representa el conjunto de las K imágenes más similares a la consulta q en función de la similitud textual, y $I\{A\}$ será igual a 1 si A es verdadero y a 0 en otro caso.

Estas técnicas de combinación multimedia son denominadas de fusión secuencial en (Díaz Galiano, 2011), donde se diferencian dos subtipos: filtrado y realimentación, que coinciden con los descritos en esta misma sección. En dicho trabajo, se define el prefiltrado como la utilización de un sistema para reducir el número de objetos donde el segundo sistema debe buscar, destacando el posible inconveniente de eliminar objetos relevantes en el primer paso.

3.3 Métodos de Fusión Multimedia

Para abordar una tarea de análisis o fusión multimedia pueden seguirse diferentes métodos o estrategias, los cuales se pueden clasificar en tres grandes grupos (Atrey, Hossain and Kankanhalli, 2010): 1) basados en reglas, 2) basados en clasificación, y 3) basados en estimación, tal y como muestra la Figura 3.3:

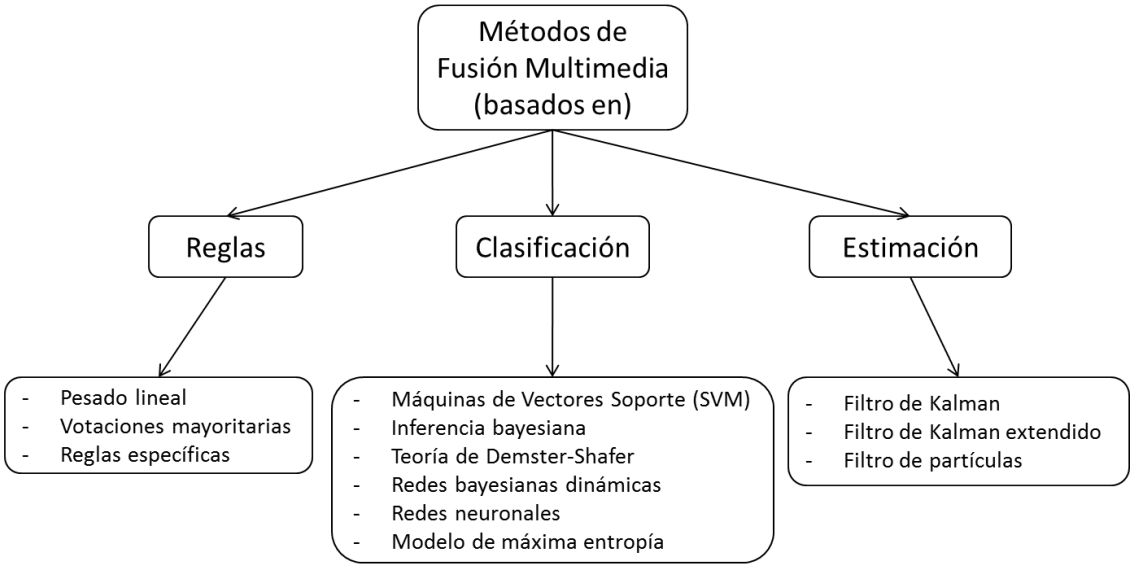

Figura 3.3. Clasificación de los métodos de fusión multimedia

A continuación se describe brevemente cada una de estas clases.

3.3.1 Métodos basados en reglas

Consistentes en definir reglas básicas para combinar la información multimodal, entre los que destacan los métodos estadísticos basados en reglas como la fusión basada en pesado lineal y en votaciones mayoritarias. Como complemento a este tipo de reglas, es posible construir reglas específicas o personalizadas. Todas estas aproximaciones funcionan correctamente si la sincronía entre los diferentes medios es alta.

Pesado lineal. Se trata de uno de los métodos más extendidos. La información obtenida de diferentes medios es combinada linealmente utilizando pesos normalizados (Jain, Nandakumar and Ross, 2005). Esta información puede provenir de características de bajo nivel como el color o el movimiento (Wang et al., 2003), o decisiones de nivel semántico, como por ejemplo las posibles ocurrencias de un determinado evento (Nock, Iyengar and Neti, 2002).

La metodología general seguida para la fusión lineal está basada en combinaciones que utilizan operadores de suma o de producto con el fin de que los clasificadores puedan proporcionar una decisión de alto nivel. Si se tiene que $I_i(i = 1..n)$ es el vector de características obtenido desde la fuente de información o modalidad *i* (texto, imagen, audio, etc.) o la decisión obtenida desde el clasificador, y que $w_i(i = 1..n)$ es el peso asignado al modo o al clasificador *i* correspondiente, se tiene que el conjunto de características o la decisión final I será:

$$I = \sum_{i=1}^{n} w_i \cdot I_i \qquad \text{o} \qquad I = \prod_{i=1}^{n} I_i^{w_i}$$

Estos métodos son computacionalmente bastante más sencillos que otros que se verán en esta sección, aunque cualquier sistema de fusión que los utilice deberá determinar y ajustar los pesos adecuadamente en función de la tarea que se esté abordando.

Cabe destacar los trabajos desarrollados al nivel de decisión para la recuperación de imágenes y fragmentos de vídeos presentes en (Hua and Zhang, 2005) y (Donald and Smeaton, 2005). Estos últimos utilizan una estrategia de fusión lineal ponderada para combinar los *scores* y *ranks* (posiciones) normalizados de los resultados en el campo de la recuperación de vídeos, en base a diferentes modalidades de información como el texto o las características visuales

(color, textura, etc.), llegando a la conclusión de que la combinación de *scores* con diferentes pesos es la mejor solución para fusionar resultados textuales y visuales.

Votación por mayoría. Se trata de un caso especial de combinación basada en pesos, en el que todos son iguales. En este caso, la decisión final es aquella donde la mayoría de los clasificadores obtienen un resultado similar (Sanderson and Paliwal, 2004). Por ejemplo, en (Radova and Psutka, 1997) se presenta un sistema de identificación de locutores que utiliza múltiples clasificadores, en el que las muestras del discurso de cada locutor serán las características. A partir de estas muestras se identifican un conjunto de patrones, que serán clasificados por dos clasificadores. Los *scores* de salida de estos clasificadores son fusionados mediante una aproximación tardía para obtener la decisión mayoritaria.

Reglas específicas. Consistente en definir reglas personalizadas que van más allá de las reglas estadísticas estándar. Ejemplo de esta aproximación puede encontrarse en (Pfleger, 2004) o (Holzapfel, Nickel and Stiefelhagen, 2004), donde se muestra una aproximación de fusión multimodal utilizando reglas específicas para combinar el habla con los gestos durante la interacción de un robot en una cocina.

3.3.2 Métodos basados en clasificación

Esta categoría incluye un conjunto de técnicas que son utilizadas para clasificar las observaciones multimodales en una clase predefinida, destacando el método *Support Vector Machine* (SVM), la inferencia Bayesiana, la teoría Dempster-Shafer, las redes bayesianas dinámicas y el modelo de máxima entropía.

Support Vector Machine (SVM). En el dominio multimedia, este método está siendo utilizado para multitud de tareas entre las que destacan la clasificación de características, la clasificación de conceptos, la detección de caras, la clasificación de texto, etc. SVM se considera un método de aprendizaje supervisado y se aplica como un clasificador lineal binario óptimo donde el conjunto de vectores de entrada se reparte entre una de los dos clases aprendidas. Desde el punto de vista de la fusión multimodal, SVM se utiliza para la clasificación de patrones. La extensión del método básico SVM permite la creación de clasificadores no lineales.

Esta técnica ha sido utilizada en multitud de tareas como la detección de conceptos semánticos en vídeos utilizando información de audio, vídeo y texto (Adams et al., 2003), el análisis semántico de vídeos (Snoek, Worring and Smeulders, 2005), o la clasificación de imágenes (Zhu, Yeh and Cheng, 2006). En este último trabajo se presenta un proceso de dos pasos. Primero, se aplica un modelo de bolsa de palabras (*BOW, bag of words*) para clasificar la imagen correspondiente considerando las características de bajo nivel. De forma paralela, el detector de texto busca la presencia del mismo en la imagen. En el segundo paso, se utiliza un clasificador SVM por pares para fusionar las características textuales y visuales. El proceso completo puede verse en la siguiente figura:

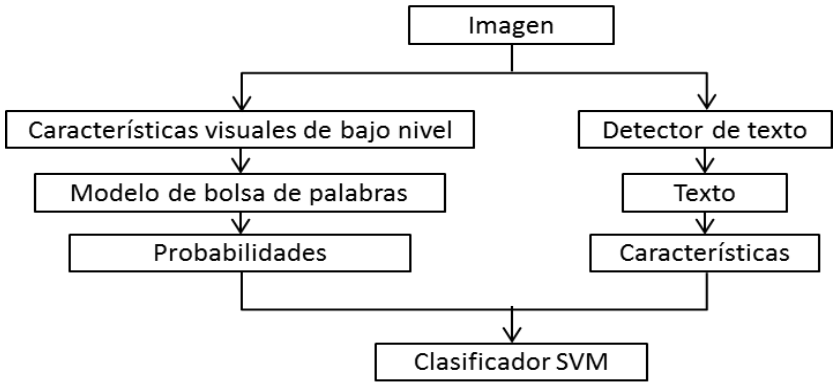

Figura 3.4. Fusión multimodal basada en SVM

Dentro de la tarea de anotación de imágenes fotográficas, en (Znaidia et al., 2012) se utiliza un clasificador lineal SVM para las características obtenidas desde las modalidades textual y visual. La combinación de los clasificadores de cada modo permite incrementar el rendimiento de los clasificadores individuales.

Inferencia bayesiana. Este método lleva a cabo la combinación de la información multimedia a partir de reglas basadas en la teoría de la probabilidad y puede ser aplicado tanto a nivel de características como a nivel de decisión. Infiere la probabilidad de una observación o una decisión a partir de la combinación de las observaciones extraídas de múltiples medios o las decisiones obtenidas a partir de diferentes clasificadores. Se trata de un método clásico, ampliamente utilizado y que ha sido extendido en otras aproximaciones.

El método de fusión basado en inferencia bayesiana se describe de la siguiente manera: se tienen los vectores de características o de decisiones (I_1, I_2,..., I_n) obtenidos a partir de las

distintas modalidades disponibles (un total de *n*). Si se asume que estas modalidades son estadísticamente independientes, la probabilidad conjunta de la hipótesis H basada en los vectores de características o decisiones fusionados será:

$$p(H|I_1, I_2, \ldots, I_n) = \frac{1}{N}\prod_{k=1}^{n} p(I_k|H)^{w_k}$$

donde N se utiliza para normalizar, y el término w_j será el peso de la modalidad *k* ($\sum_{k=1}^{n} w_k = 1$).

El método de inferencia bayesiana ha sido utilizado en la fusión de información multimodal para abordar diferentes tareas de análisis como el reconocimiento audiovisual del habla, el análisis de vídeos deportivos, o la detección de eventos en el dominio de la vigilancia multimedia.

Aunque el método de fusión basado en inferencia bayesiana permite modelar incertidumbre, algunos investigadores prefieren el uso de la teoría Dempster-Shafer (Reddy, 2007), que generaliza la teoría Bayesiana permitiendo asignar evidencia a la unión de hipótesis.

Redes bayesianas dinámicas. La inferencia Bayesiana puede extenderse a una red o grafo donde cada nodo represente variables aleatorias (observaciones o estados) de diferentes tipos (por ejemplo, audio o vídeo) y los enlaces denoten las dependencias probabilísticas. Una red bayesiana se comporta dinámicamente cuando incorpora aspectos temporales. Dado que estos modelos describen los datos observados en términos del proceso que los ha generado reciben también el nombre de generativos. Sus ventajas frente a las aproximaciones anteriores se basan en su capacidad de modelar múltiples dependencias entre los nodos y su facilidad para integrar información temporal en el modelo, lo que las hace muy adecuadas en tareas de análisis donde sea necesario tomar decisiones que impliquen evoluciones temporales.

Los modelos ocultos de Markov (HMM, *hidden Markov models*) pueden ser considerados como una red bayesiana dinámica. Un ejemplo de su utilización puede verse en (Chaisorn et al., 2003), donde se utilizan características multimodales (textuales y audiovisuales) en una tarea de segmentación de vídeos de noticias en historias individuales.

En (Wu, Chang and Tsengh, 2005) se sigue una aproximación basada en un diagrama de influencia (un tipo de red bayesiana) con el objetivo de representar la semántica de las imágenes. Este marco de fusión multimedia integra información contextual (localización, tiempo, y parámetros de cámara) e información de contenido (holística y localizada), con la ontología semántica orientada al dominio, representada por un grafo dirigido.

Redes Neuronales. Las redes neuronales son otra aproximación para la fusión de información multimedia. Pueden entenderse como una caja negra que, una vez entrenada, es capaz de resolver problemas computacionalmente muy costosos (Wu, 2003). El método consiste en crear una red con tres tipos de nodos (entrada, oculto, salida). Los nodos de entrada reciben observaciones o decisiones y los nodos de salida proporcionan los resultados de la fusión de los nodos de entrada. En este método, el diseño de la red, así como los pesos asignados a los diferentes caminos dentro de ésta resultan críticos para su éxito. Estos pesos pueden ser ajustados en la fase de entrenamiento con el fin de obtener los resultados óptimos. Su aplicación puede hacerse tanto al nivel de características como al nivel de decisiones.

Modelo de máxima entropía. Se trata de un clasificador estadístico que sigue una aproximación basada en la teoría de la información y que determina la probabilidad de pertenencia de una observación a una determinada clase, en función de su propio contenido.

Un ejemplo de aplicación de este modelo puede verse en (Magalhaes and Ruger, 2007), donde se aborda la tarea de indexación semántica multimedia. En dicho trabajo se combinan el texto y las características visuales de la imagen para la posterior recuperación.

3.3.3 Métodos basados en estimación

Esta categoría incluye los métodos de fusión basados en el filtro de Kalman (y su versión extendida) y el filtro de partículas. El filtro de Kalman procesa en tiempo real información dinámica de bajo nivel proporcionando estimaciones estadísticas de estado a partir de la información fusionada.

El filtro de partículas se refiere a un conjunto de métodos de simulación utilizados para estimar la distribución de estados en un modelo no lineal y no gaussiano. Se conocen también como métodos secuenciales Monte Carlo. Han sido utilizados en análisis multimedia para

hacer el seguimiento de un locutor a partir de información audiovisual o para hacer el seguimiento de múltiples personas en el entorno de una videoconferencia.

3.4 Algoritmos de Fusión Tardía

Como se ha indicado en el apartado 3.2.2, estos algoritmos permiten combinar los resultados o decisiones obtenidas individualmente por cada uno de los sistemas monomodales (TBIR y CBIR en el caso de la recuperación multimedia de imágenes). Cada uno de estos módulos generará su propia lista de resultados con las imágenes recuperadas ordenadas en función de valor (*score*) de relevancia o similitud obtenido para cada consulta o *topic*. Estos resultados monomodales son posteriormente combinados para obtener la lista multimedia final.

3.4.1 Funciones de agregación de *scores*

Para fusionar listas de resultados siguiendo una aproximación basada en fusión tardía se utiliza una función de agregación de sus *scores*, también llamadas reglas de combinación en (Zhou, Depeursinge and Müller, 2010).

Formalmente, este tipo de funciones de agregación pueden definirse de la siguiente manera (para el caso de combinar dos listas de resultados, una textual y otra visual):

$$S_{late}(q,d) = nz(d)^{y}\left(\alpha_t N\big(s_t(q,d)\big) + \alpha_v N\big(s_v(q,d)\big)\right)$$

donde:

- $nz(d)$: número de modalidades (sistemas monomodales de recuperación) para los que el documento *d* es recuperado (*score*>0, *non zero*) como resultado para la consulta *q*.

- *y*: parámetro utilizado para asignar la importancia al número de sistemas monomodales que recuperan un determinado resultado. Cuando *y*=0 la función de agregación se convertirá en la media aritmética. Con *y*=1 se estará agregando en base al algoritmo de fusión tardía *combMNZ*, que se explicará con más detalle más adelante.

- α_t, α_v: pesos (o valores de confianza) dados a cada uno de los modos. La suma de ambos pesos ha de ser igual a 1.

- N: operador de normalización. Transforma un conjunto de *scores* de similitud, o valores de relevancia, en valores dentro del rango 0-1. En el apartado 3.6 se describirán las técnicas de normalización más utilizadas.

- $s_t(q,d)/s_v(q,d)$: *score* obtenido desde el sistema de recuperación textual (TBIR) / visual (CBIR) por el objeto multimedia d para la consulta q.

A partir de la fórmula mostrada pueden llevarse a cabo diferentes tipos de agregación en función de los valores dados a los parámetros, como son (Fox and Shaw, 1994), (Lee, 1997):

- *combMAX*: el *score* resultante tras la fusión para el objeto multimedia *d* en relación a la consulta *q* será el mayor de entre todos los *scores* obtenidos desde los *N* distintos sistemas monomodales (s_t y s_v para el caso de la recuperación multimedia de imágenes).

$$s_{combMAX}(q,d) = max_{k=1..N}(s_k(q,d))$$

  Este algoritmo no es muy robusto a errores, ya que se basa en solo una de las listas monomodales para cada imagen resultado.

- *combMIN*: esta combinación consistirá en seleccionar el mínimo *score* de entre las *N* listas de resultados, lo contrario a *combMAX*.

$$s_{combMIN}(q,d) = min_{k=1..N}(s_k(q,d))$$

- *combSUM*: combina los resultados de las diferentes listas monomodales calculando la suma de todos los *scores* obtenidos por el objeto multimedia *d* para la consulta *q*.

$$s_{combSUM}(q,d) = \sum_{k=1}^{N} s_k(q,d)$$

  La principal desventaja de este algoritmo es que el *score* final fusionado pondera por igual a todos los sistemas monomodales, incluidos aquellos con peores resultados. Para tratar de evitar este problema, existe la variante *combSUM(n)MAX*, en el que únicamente se suman los N *scores* con mayor valor,

con lo que la estabilidad de la función de combinación aumenta, ya que se evita contabilizar *scores* procedentes de sistemas monomodales con rendimientos bajos.

- *combMNZ*: esta combinación de *scores* es una variante de *combSUM* que da más importancia a los documentos recuperados desde varios modos (aparecen en varias listas de resultados). El número de sistemas monomodales que recuperan el documento *d* para la consulta *q* es indicado por la función $F(q,d)$.

$$s_{combMNZ}(q,d) = F(q,d)\sum_{k=1}^{N} s_k(q,d)$$

- *combANZ*: consiste en otra variante de *combSUM*, pero en este caso dividiendo la suma de todos los *scores* monomodales entre el número de sistemas que recuperan el objeto *d* para la consulta *q*.

$$s_{combANZ}(q,d) = \frac{1}{F(q,d)}\sum_{k=1}^{N} s_k(q,d)$$

- *combGMNZ*: generaliza las funciones combSUM y combMNZ para poder manejar (mediante el parámetro $y \geq 0$) la forma en que a los objetos multimedia recuperados por más sistemas monomodales se le otorga más peso (confianza) que a los demás. Cuando $y = 0$ la combinación sería equivalente *combSUM*, y para $y = 1$ equivalente a *combMNZ*.

$$s_{combGMNZ}(q,d) = F(q,d)^{y}\sum_{k=1}^{N} s_k(q,d)$$

Según (Fox and Shaw, 1994), (Lee, 1997) y (Renda and Straccia, 2003), *combMNZ* está considerado como el mejor método de fusión de *rankings* o listas de resultados, ya que está basado en el hecho de que diferentes motores de búsqueda recuperan conjuntos similares de documentos relevantes pero diferentes de no relevantes (Lee, 1997), y este método premia a los documentos en común de las distintas fuentes monomodales.

Diversos experimentos desarrollados indican que este tipo de fusión de listas de resultados basada en *scores* funciona mejor que la fusión basada en las posiciones o *ranks* de los documentos multimedia recuperados en las listas de resultados (Renda and Straccia, 2003).

3.4.2 Combinación basada en sistemas de votación

Estas técnicas de fusión de resultados están basadas en los sistemas de votación clásicos, aprovechando la analogía existente entre dichos sistemas de votación y la combinación o fusión de *rankings* de distintos sistemas de recuperación (metabuscadores). Los dos algoritmos más representativos de este tipo de técnicas de combinación de listas de resultados son:

- *Borda-fuse* (Aslam and Montague, 2001): modelo de fusión basado en el sistema de votación óptima *Borda Count*. Se trata de un método de votación basado en las posiciones obtenidas por los objetos recuperados en las listas de resultados (*rank*), que aplica una penalización lineal determinada por dicho *rank*. Se trata de un método de fusión de listas de resultados computacionalmente muy sencillo, ya que se puede implementar con complejidad lineal. Define un sistema electoral basado en el consenso, y no en la mayoría.

  El funcionamiento del algoritmo consiste en asignar a cada objeto de la lista de resultados un valor en función del *rank* que ocupa. Si el número de elementos de una lista de resultados es N, el número de puntos otorgados al primer clasificado será N, al segundo N-1, y así sucesivamente. Es a lo que se llama normalización *Borda*. Las puntuaciones obtenidas por cada objeto en cada lista se suman y, de esta manera, se generará la lista de resultados final fusionada. La combinación basada en *Borda-fuse* es equivalente al algoritmo *combSUM* una vez que se ha realizado la normalización *Borda*.

  Un problema que muestra esta técnica es que si para una determinada consulta un sistema devuelve una lista de únicamente 2 resultados, según la normalización *Borda*, el primer objeto del *ranking* obtendrá solo 2 puntos, que serán los mismos que obtendría un objeto recuperado en la posición 999 en una lista de 1000 resultados.

Existe una variante de este método, el *Borda-fuse* ponderado (*Weighted Borda-fuse*, (Aslam and Montague, 2001)), en el que se da peso a cada sistema de entrada, y sus puntos se multiplican por dicho peso.

- *Condorcet-fuse* (Montague and Aslam, 2002): realiza una fusión basada en el método de votación *Condorcet*, consistente en seleccionar al candidato que gane por mayoría (o empate) en todos los emparejamientos contra el resto de candidatos (si es que alguno de los candidatos cumpliese esa propiedad).

3.5 Fusión multimedia en la recuperación de imágenes

En la tarea de recuperación de imágenes, la búsqueda basada en texto es más eficiente y precisa que la visual (Clinchant, Csurka and Ah-Pine, 2011), aunque una adecuada combinación de la información disponible desde ambos modos podría resultar beneficiosa (Chatzichristofis et al., 2010), si se explotasen las ventajas de cada modalidad y la complementariedad existente entre ellas.

La mayor parte de técnicas de fusión multimedia combinan, mediante funciones de agregación (descritas en el apartado 3.4.1), los *scores* obtenidos tras los procesos de recuperación monomodal (TBIR y CBIR), ya que estos algoritmos basados en fusión tardía (*late fusion*) funcionan mejor que los basados en fusión temprana (Depeursinge and Muller, 2010). Adaptando estos algoritmos de agregación, como *combMAX*, *combSUM* y *combMNZ* (Fox and Shaw, 1994), a la tarea concreta de la recuperación multimedia de imágenes, el primero de ellos calcularía el *score* fusionado como el valor máximo obtenido por una determinada imagen *i* en las listas monomodales de resultados a combinar: lista de resultados textuales (*score* S_t) y visuales (S_v), esto es, $max(S_t, S_v)$. En el caso de utilizar *combSUM*, se calculará el *score* combinado como la suma de los *scores* monomodales ($S_t + S_v$). En el caso de *combMNZ* se otorgará más importancia a las imágenes recuperadas por ambos subsistemas. Un ejemplo de aplicación del algoritmo *combSUM* utilizado en el marco de la tarea de recuperación de imágenes de dominios médicos puede encontrarse en (Vanegas et al., 2012), donde se utiliza para fusionar los resultados recuperados en base a cada una de las imágenes de ejemplo proporcionadas como consulta.

Analizando las soluciones propuestas y los resultados obtenidos por los diferentes grupos participantes en la tarea de recuperación multimedia de imágenes de Wikipedia organizada

dentro del foro de evaluación *ImageCLEF* (en el que el autor de esta memoria ha participado desde la edición de 2008 hasta la de 2011), se observa que la fusión multimedia basada en las características textuales y visuales de las imágenes no ha sido capaz de mejorar los resultados monomodales de los sistemas TBIR hasta hace relativamente poco. Durante las ediciones de 2008 y 2009 los sistemas basados únicamente en texto resultaban imbatibles para las aproximaciones basadas en fusión multimedia (Tsikrika, 2010), consiguiendo mejores resultados únicamente para ciertas consultas. Es en la edición de 2010 cuando por primera vez, y únicamente en el caso de dos de los grupos participantes, los mejores sistemas de recuperación basados en fusión multimedia mejoran a las aproximaciones textuales (Tsikrika, 2011). De nuevo en la edición del 2011 el mejor sistema está basado en un enfoque multimedia, pero en este caso esto sucede para la mayoría de los grupos participantes: 8 de 9 obtienen sus mejores resultados con sistemas multimedia (Tsikrika, Popescu and Kludas, 2011).

A lo largo de las distintas ediciones del foro de evaluación ***ImageCLEF*** (la primera fue en el año 2003) se han observado principalmente tres tipos de aproximaciones: 1) basadas en fusión temprana (*early fusion*), 2) realimentación entre modos (textual y visual) mediante expansión de la consulta (QE, *query expansion*), y 3) fusión a nivel de decisiones o tardía (*late fusion*), que ha sido la aproximación más utilizada con diferencia. La siguiente figura muestra esta categorización (Muller, Clough and Desealaers, 2010).

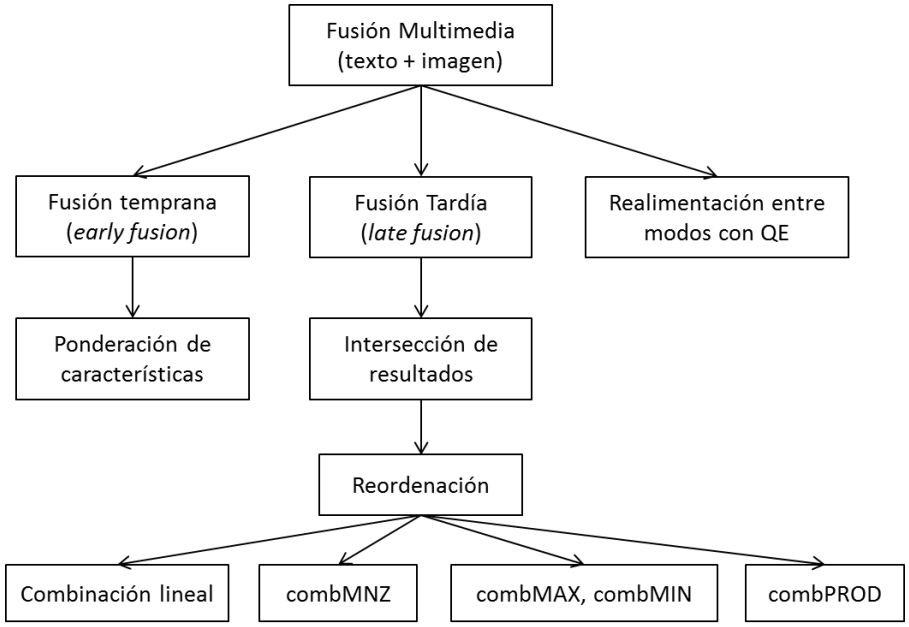

Figura 3.5. Técnicas de fusión utilizadas en *ImageCLEF*

La **fusión temprana** o a nivel de características, consistente en combinar la información procedente de los modos textual y visual antes de tomar ninguna decisión, conlleva la dificultad relacionada con la gran heterogeneidad de ambos espacios de características. La combinación entre variables que son términos discretos (términos textuales), con otras que son continuas como las características visuales, no es algo trivial y en algunos casos se producen interacciones negativas (Bell, 2003). Un buen resultado de este tipo de técnicas se da en la tarea de recuperación de imágenes médicas de la edición de 2009, donde el mejor rendimiento lo obtuvo un trabajo basado en una aproximación de este tipo (Berber and Alpkocak, 2009). En él se combinaban las características textuales con un pequeño número de descriptores visuales.

Otro ejemplo de fusión temprana con mejora con respecto a una aproximación basada en fusión tardía (utilizando la regla *combMIN*), aunque sin significancia estadística, puede verse en (Ferecatu and Sahbi, 2008) donde las características textuales y visuales son normalizadas antes de ser concatenadas, esto es, no se utiliza ningún tipo de pesado para las características. Por el contrario, en (Moulin et al., 2009) (Salton and Buckley, 1988) se observa una degradación del rendimiento de la recuperación trabajando con la colección de imágenes de Wikipedia. En este trabajo se crea inicialmente un vocabulario visual a partir de los descriptores básicos de las imágenes, lo que posteriormente se fusiona con las características

textuales utilizando un pesado *tf-idf* (Robertson and Jones, 1976). La fusión a nivel de características es también utilizada para la tarea de recuperación de imágenes médicas en (Seco, Markonis and Muller, 2012), donde combinan los distintos descriptores visuales de cada una de las imágenes de consulta en un único vector antes de calcular la distancia (similitud) con el resto de imágenes de la colección.

La solución basada en utilizar la información disponible en uno de los modos para **expandir** las consultas antes de realizar la búsqueda en base al otro modo, tiene inicialmente la intención de mejorar el nivel de cobertura de sus resultados, ya que los términos textuales o imágenes de ejemplo adicionales permitirán, en teoría, recuperar más imágenes relevantes (pseudo-relevancia). El problema de esta técnica es que, a la vez, se corre el riesgo de recuperar demasiados resultados y disminuir la precisión de los resultados. Un claro ejemplo de este tipo de aproximación puede verse en (El Demerdash, Kosseim and Bergler, 2009). En cuanto a la dirección de la expansión en mucho más común encontrar técnicas que modifican la consulta textual añadiéndole nuevos términos extraídos de las primeras imágenes recuperadas visualmente. También se observa la alternativa de expandir la consulta textual haciendo uso de conceptos detectados en las imágenes de ejemplo de la consulta multimedia, como en (Popescu, Le Borgne and Moellic, 2008) o (Inoue and Grover, 2008). Otro ejemplo de expansión de consulta con información visual realizado sobre los resultados textuales es utilizado en (Daroczy, Petras and Benczur, 2010), obteniendo una ligera mejora en los resultados finales.

Utilizar la información de uno de los modos (visual o textual) para reordenar (*re-ranking*) los resultados obtenidos por la recuperación basada únicamente en el otro es otra de las técnicas comúnmente utilizadas. En (Daroczy, Pethes and Benczur, 2011) se utiliza la información visual para reordenar los resultados obtenidos desde el sistema de recuperación basada en texto, lo que se conoce como *image re-ranking*. El rendimiento del sistema resulta ligeramente peor en comparación con la recuperación puramente textual.

La aproximación de fusión multimedia más utilizada por los grupos participantes en las distintas ediciones de *ImageCLEF* ha sido, con diferencia, la basada en técnicas de fusión a nivel de decisiones (***late fusion***). Este tipo de fusión permite combinar los resultados de cualquier tipo de sistema de recuperación que devuelva una lista de imágenes. Por esto, cuando se consiguen unos resultados monomodales aceptables, los esfuerzos suelen centrarse

en encontrar una técnica de fusión que explote la complementariedad entre la información proporcionada por ambos sistemas (Zhou, Depeursinge and Müller, 2010). La variedad encontrada entre las diferentes posibilidades de estrategias de fusión tardía es bastante alta. Pueden encontrarse estrategias de fusión que basen su algoritmo en el *score* (valor de relevancia o similitud), o que lo hagan en el *rank* (posición en la lista de resultados). Por ejemplo, en (Arampatzis, Chatzichristofis and Zagoris, 2010) llevan a cabo experimentos basados en la combinación de *scores*, llegando a la conclusión de la mayor importancia de la modalidad textual en relación a la visual, dada la escasa mejora apreciada cuando se incorpora esta al esquema global de recuperación. Otra opción para la combinación consiste en realizar la intersección entre las distintas listas de resultados, como por ejemplo en (Villena-Roman et al., 2007), donde se definen varios operadores de combinación. Otra de las soluciones más utilizadas en *ImageCLEF* consiste en reordenar las imágenes recuperadas textualmente en base a sus *scores* visuales, como por ejemplo en (Zhou, Gobeill and Muller, 2008), (Mulhem et al., 2009) o (Simpson et al., 2009). La aproximación inversa, esto es, reordenar en base a los *scores* textuales las imágenes recuperadas por el sistema visual también está presente en trabajos como (Gao and Lim, 2009) o (Hare, Dupplaw and Lewis, 2009).

Siguiendo el repaso de las técnicas de fusión tardía utilizadas en el foro de evaluación ImageCLEF, hay que destacar las basadas en combinaciones lineales, que generan la lista de resultados fusionada en base a los *scores* tanto textuales (S_t) como visuales (S_v) obtenidos. Esta combinación lineal entre *scores* puede representarse de la siguiente manera:

$$S_{fusion}(d) = \alpha \cdot S_t(d) + (1 - \alpha) \cdot S_v(d)$$

Habitualmente, el valor de ponderación α ha sido definido con valores altos para otorgar más confianza a la parte textual, basándose en los mejores resultados monomodales obtenidos por dicha modalidad. Una excepción a esta regla puede encontrarse en (Douze et al., 2009), donde se obtienen mejores resultados cuando se proporciona un peso alto al *score* visual. Algunas aproximaciones tratan de calcular los pesos a partir de la información disponible de ediciones anteriores, como (Ruiz, 2009). Un caso interesante está en (Liu et al., 2012), donde se utiliza una suma ponderada para fusionar los *scores* provenientes de texto y de imagen, aprendiendo los pesos de ponderación en una fase previa de entrenamiento. Este tipo de fusión tardía es llamada SWLF (*Selective Weighted Late Fusion*) por los autores.

Entre los algoritmos concretos de combinación utilizados, cabe destacar varios. *combSUM* sería equivalente a utilizar valores de pesado α = 0.5 para otorgar la misma confianza a ambas modalidades (Navarro, Muñoz and Llopis, 2009). Una variante a este, conocido como *combMNZ*, fue ligeramente modificado en (Inkpen et al., 2008) para la tarea de recuperación de imágenes fotográficas, aplicando un peso a los *scores* normalizados de cada modalidad para controlar sus respectivas influencias. Los métodos que focalizan toda su confianza en una de las modalidades, según la consulta, como son *combMAX* y *combMIN* también han sido utilizados en las tareas de recuperación de imágenes fotográficas y médicas. Una solución que combina estas dos medidas aparece en (Villena-Roman et al., 2007) donde el *score* tras la fusión es:

$$S_{fusion}(d) = combMAX + \frac{combMIN^2}{combMAX + combMIN}$$

Otro ejemplo de combinación lineal entre resultados textuales y visuales es llevada a cabo en (Boros, Ginsca and Iftene, 2011) donde, tras múltiples experimentos, se define la combinación en base al cálculo de la media ponderada entre los *scores* textuales obtenidos desde Lucene (S_t), y los valores de similitud visuales normalizados (S_v). Con este tipo de fusión, los autores obtienen sus mejores resultados cuando el peso de la componente textual supone un 60% del total.

En (Cao et al., 2012), donde se aborda una tarea de clasificación de imágenes sobre un conjunto de conceptos médicos, inicialmente se clasifican las imágenes en base a cada grupo de características visuales y, posteriormente, se combinan los resultados de cada clasificación. También en (Yan et al., 2012), puede observarse una aplicación de fusión tardía en la que se combinan resultados procedentes únicamente de la recuperación basada en texto (no multimedia), con el objetivo de anotar conceptos semánticos dentro de una colección de imágenes de *Flickr*. Se fusionan los *scores* normalizados de las listas de imágenes resultado obtenidas a partir de dos campos de información textual diferentes (*title* y *image*), siguiendo una combinación lineal que otorga el mismo a ambas listas.

Las combinaciones lineales basadas en la posición o *rank* obtenido por las imágenes en las listas de resultados son bastante menos utilizadas y, aunque tienen la ventaja de no necesitar

ningún tipo de normalización, pierden el valor de confianza de las modalidades al fijarse únicamente en la posición, sin tener en cuenta su valor de semejanza.

3.6 Técnicas de Normalización

La normalización consiste en hacer que el rango de los *scores* calculados para los objetos recuperados desde diferentes sistemas de recuperación, sean comparables entre ellos (Wu, Crestani and Bi, 2006). Esta necesidad surge debido a que los distintos sistemas utilizan estrategias muy diferentes para calcular y asignar el score de relevancia a los documentos recuperados. Esta diferencia puede estar relacionada tanto con el rango de los valores como con su distribución. El rendimiento del sistema utilizando *scores* normalizados para la fusión depende en gran medida de la propia definición de dicho *score* en cada experimento (Zhou, Depeursinge and Müller, 2010).

Los métodos de normalización propuestos en la literatura pueden clasificarse en base a si están basados en la posición que ocupan los objetos recuperados dentro de una lista de resultados (*rank*), o en el valor de relevancia o similitud obtenido por cada uno de ellos (*score*). Algunos trabajos como (Wu, Crestani and Bi, 2006) o (Aslam and Montague, 2001) han confirmado que los *rankings* no son tan informativos como los *scores* cuando se abordan tareas de fusión de información.

En (Zhou, Depeursinge and Müller, 2010) se hace referencia a algunas de las técnicas de normalización más utilizadas. Una de ellas es la propuesta en (Lee, 1997), conocida como normalización *min-max*, que calcula el nuevo *score* de relevancia en función de los *scores* máximo (s_{max}) y mínimo (s_{min}) obtenidos:

$$Norm(s) = \frac{s - s_{min}}{s_{max} - s_{min}}$$

La ecuación anterior calculará los nuevos *scores* de relevancia normalizados para cada una de las imágenes recuperadas. El objeto recuperado con el máximo *score* antes de la normalización, tendrá a posteriori un *score* igual a 1. Para el caso del objeto con el score mínimo, su nuevo *score* tras el proceso de normalización será 0. Según se ha definido esta técnica de normalización, se generarán dos grupos de *scores* normalizados: 1) cuando se aplica el método a nivel de experimento o *run*, y 2) cuando se aplica a nivel de *topic* o

consulta. Esta técnica de normalización también es conocida como método *Zero-one* según (Wu, Crestani and Bi, 2006), donde también se propone la variante *fitting method* (método de prueba) que utiliza un rango de valores menor [a, b] con $(0 < a < b < 1)$, ya que en algunos casos los documentos recuperados en las primeras posiciones no son siempre relevantes y, de igual modo, los recuperados al final de la lista pueden no ser irrelevantes.

Otra estrategia consiste en convertir el *ranking* de cada resultado en un valor de similitud (*score*). De este modo se evita el denominado "efecto de independencia del pesado", que tiende a tener en cuenta los resultados individuales de cada *run* sin considerar su rendimiento global. La siguiente fórmula muestra cómo convertir la posición (*rank*) del objeto recuperado en un valor de relevancia o *score*:

$$Score(rank) = \frac{rank - 1}{N_{recuperados}}$$

La normalización lineal se calcula en función de los *ranks* de cada objeto multimedia recuperado, que penaliza los *ranks* más bajos al ser menos relevantes:

$$Norm_{lineal}(rank) = N_{recuperados} - rank$$

Este tipo de normalización tendrá sentido cuando todas las listas de resultados tengan el mismo número de resultados, con *rank* entre 1 y $N_{recuperados}$. Diferentes experimentos han demostrado que para la mayoría de sistemas de recuperación de información, el rendimiento tiende a decrecer de manera logarítmica (Vogt and Cottrell, 1999). Como consecuencia de esto, en (Zhou, Depeursinge and Müller, 2010) se propone una función de normalización con penalización logarítmica:

$$Norm_{log}(rank) = \ln\left(N_{recuperados}\right) - \ln(rank)$$

Para (Zhou, Depeursinge and Müller, 2010), esta última técnica de normalización es la que mejores resultados obtiene, y la que resulta ser más estable.

En (Montague and Aslam, 2001) se propone el método *Sum*, consistente en transformar el *score* mínimo en 0 y la suma de todos los *scores* en 1, y el método *ZMUV* (*Zero-Mean and Unit-Variance*) o *z-score*, donde la media de todos los scores se transforma en 0, y su varianza en 1. El método de prueba es más favorable cuando se utiliza la fusión *CombSum*, mientras

que *Zero-one* es el mejor con *CombMNZ*. Por el contrario, el método *Sum* no es tan bueno como los dos primeros (aunque los mejora en algún caso) y *ZMUV* siempre obtiene los peores resultados y no parece ser un método de normalización adecuado.

Otro posible método de normalización es definido en (Renda and Straccia, 2003). Se trata de la normalización *Borda* basada en el *rank*:

$$Norm_{Borda}(rank) = 1 - \frac{rank - 1}{N_{total}} \qquad SI\ objeto \in lista\_recuperados$$

$$Norm_{Borda}(rank) = \frac{1}{2} + \frac{N_{recuperados} - 1}{2 \cdot N_{total}} \qquad e.o.c.$$

Según la ecuación, el método consistirá en asignar puntos a cada objeto en función de la posición (*rank*) en la que aparece en la lista de resultados. Para el caso de la tarea de combinación de listas en metabuscadores, el método de normalización *Score* propuesto por (Lee, 1997) funciona mejor que la normalización *Borda* del *rank* (Renda and Straccia, 2003).

En el apartado 7.3.4 se llevará a cabo un análisis referente a la inclusión de una fase de normalización previa a la combinación de resultados monomodales dentro de la estrategia del enfoque de fusión multimedia propuesto en esta tesis. Se analizarán diversos experimentos que evaluarán la conveniencia o no de llevar a cabo la normalización de las listas de resultados.

Capítulo 4 Marco de Evaluación

Se define inicialmente la metodología de evaluación seguida, y se describe el foro de evaluación *ImageCLEF* donde el autor de esta memoria ha participado desde la edición de 2008 hasta la de 2013 en diferentes tareas relacionadas con la recuperación multimedia de imágenes. A continuación se describen las colecciones utilizadas, y finalmente se describen las herramientas empleadas, algunas de desarrollo e implementación propia, para las tareas de preprocesamiento textual, indexación y recuperación, algoritmos de fusión, y evaluación de resultados.

4.1 Metodología de evaluación

La tarea de evaluación de sistemas de recuperación, ya sean de información textual o de información multimedia, consiste en medir cómo de bien los sistemas satisfacen las necesidades de información de los usuarios (Baeza-Yates and Ribeiro-Neto, 2011). En base al uso de diferentes métricas podrá determinarse el rendimiento de los sistemas, y se podrán comparar varios de ellos entre sí.

Los métodos de evaluación utilizados en los más conocidos foros y congresos científicos relacionados con la recuperación de información (TREC, CLEF, NTCIR), surgen de los experimentos realizados a mediados del siglo pasado en la Universidad de Cranfield (Cleverdon, 1960). Es lo que se conoce como paradigma de “evaluación Cranfield”.

Los tres elementos fundamentales de este paradigma, para llevar a cabo la evaluación de distintos sistemas de recuperación bajo condiciones similares, son:

1) Colección de documentos (que serán objetos multimedia para el caso de los sistemas de recuperación multimedia).

2) Conjunto de consultas, que expresarán las necesidades del usuario dentro de un escenario real (para la recuperación multimedia, las consultas podrán estar formadas por texto, imagen, audio y/o vídeo).

3) Juicios de relevancia, indicando el conjunto de documentos/objetos de la colección que son considerados relevantes para cada una de las consultas o necesidades de información.

Los juicios de relevancia (*qrels* o *ground truth*), o resultados esperados para un conjunto de consultas, tendrán valor 1 si el documento se considera relevante para la consulta, y 0 en caso de considerarse no relevante. Estos juicios pueden generarse mediante la ayuda de expertos humanos en el dominio o mediante técnicas de *pooling*. El paradigma de evaluación de Cranfield asume que todos los documentos de la colección están evaluados con respecto a cada consulta, lo cual resulta complicado cuando se trabaja con colecciones grandes. Es en estos casos cuando se utilizan métodos de *pooling*, consistentes en seleccionar los resultados recuperados entre los primeros por varios sistemas de recuperación, y evaluar únicamente esos.

4.1.1 Medidas de evaluación

La evaluación de un sistema de recuperación debe realizarse tanto desde el punto de vista de la eficiencia como desde el de la efectividad (Ingwersen and Jrvelin, 2011). Por un lado, la eficiencia mide aspectos como el tiempo de cómputo o de ejecución, el tamaño de la memoria de almacenamiento, etc. Por otro, las medidas de efectividad se encargan de la calidad de los resultados, esto es, cómo de adecuados son éstos para las necesidades de información expresadas por los usuarios.

El trabajo que aquí se presenta está mayormente relacionado con el intento de mejorar el segundo tipo de evaluación, la que se centra en la efectividad de las técnicas y algoritmos propuestos para la tarea de recuperación multimedia de imágenes, aunque algunos aspectos también se relacionan con la eficiencia. Se muestran a continuación las medidas de evaluación más utilizadas cuando se trata de medir la efectividad de un sistema de recuperación.

Las dos medidas básicas para evaluar la efectividad dentro del campo de la recuperación de información son la **cobertura** (*recall*) y la **precisión**. Estas medidas se calculan en función de los aciertos y errores cometidos en la lista de resultados, para lo que será necesario conocer a priori los juicios de relevancia correspondientes a las consultas de evaluación utilizadas. En dichos juicios se dispondrá de información acerca de la relevancia o no de los documentos u objetos multimedia (por ejemplo, imágenes) con respecto a las consultas. A partir de dicha información, se calculará la cobertura como la proporción entre el número de objetos relevantes recuperados y el de objetos relevantes existentes en total. La precisión será la proporción entre el número de objetos relevantes recuperados y el de objetos recuperados en total.

$$Cobertura = \frac{|objetos\ relevantes\ recuperados|}{|objetos\ relevantes\ en\ la\ colección|}$$

$$Precisión = \frac{|objetos\ relevantes\ recuperados|}{|objetos\ recuperados|}$$

En la siguiente figura pueden observarse gráficamente los conjuntos de objetos recuperados (N), de objetos relevantes existentes en la colección (R), y de objetos relevantes recuperados, que será la intersección entre los dos primeros ($N \cap R$).

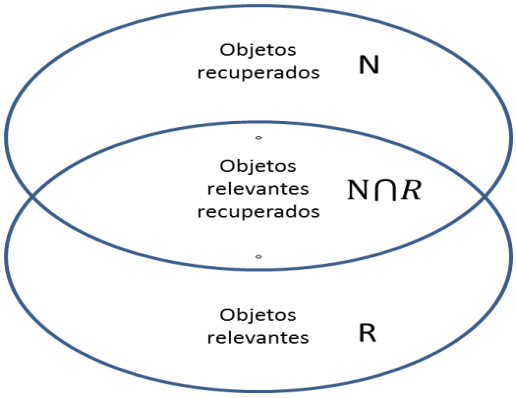

Figura 4.1. Conjuntos de documentos relevantes y recuperados

Existe la opción de combinar estas dos medidas clásicas de evaluación mediante el uso de la denominada **medida-F** (*F-measure*) o F1. Con esta medida será posible expresar la efectividad de un sistema de recuperación con un único valor. La fórmula general para esta medida es:

$$F_\beta = (1 + \beta^2) \cdot \frac{Precisión \cdot Cobertura}{(\beta^2 \cdot Precisión) + Cobertura}$$

con β real y positivo. Para el caso concreto de $\beta = 1$, se convierte la medida-F en la media armónica entre la precisión y la cobertura:

$$F_1 = 2 \cdot \frac{Precisión \cdot Cobertura}{Precisión + Cobertura}$$

Otro tipo de medida de evaluación son las precisiones calculadas hasta diferentes posiciones de la lista de resultados. Son las denominadas precisiones a bajo nivel (*early precision*) y se representan con **P@N** (*precision at N*), donde N es el número de resultados tenidos en cuenta en cada caso, esto es, la posición en la lista de resultados hasta la que se evalúa. Por ejemplo, una P@5 medirá la precisión dentro de los 5 primeros resultados recuperados por el sistema. Esta medida resulta de especial interés para los buscadores web y multimedia, ya que es conocido que en la mayoría de las ocasiones los únicos resultados en los que un usuario real está interesado son los primeros de la lista (Hearst, 2009).

La medida de evaluación **MAP** (*mean average precision*) calcula la media de las precisiones medias entre un conjunto de consultas (Voorhees, 2006). Geométricamente, es el equivalente al área bajo la gráfica no interpolada de la cobertura y la precisión. MAP se basa en mucha más información que otras medidas y es más fuerte y estable (Buckley and Voorhees, 2000).

$$MAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q}$$

donde Q es el número de consultas, y **AP** (*averge precision*) es la precisión media para la consulta q. AP es una medida que tiene en cuenta el orden en que se recuperan los documentos, no como la precisión y la cobertura. Esta precisión media (AP) se calcula en base a la siguiente ecuación:

$$AP = \frac{\sum_{k=1}^{n} (P(k) \cdot rel(k))}{|objetos\ relevantes|}$$

donde $P(k)$es la precisión a nivel k (P@k), y $rel(k)$ indica si el objeto en la posición k es relevante (valor 1) o no (valor 0).

Otras medidas de evaluación utilizadas en el foro *ImageCLEF* son Rprec, *bpref*, GMAP y MIAP. **Rprec** hace referencia a la precisión a nivel R, esto es, calcula la precisión obtenida hasta los primeros R objetos recuperados para cada consulta (siendo R el número de objetos relevantes para cada consulta).

bpref es una medida de evaluación diseñada para casos en los que los juicios de relevancia no son completos. Calcula una relación de preferencia de cuando los objetos relevantes son recuperados por delante de los no relevantes. Por lo tanto, está basada únicamente en la posición relativa de los objetos. Se define como:

$$bpref = \frac{1}{R}\sum_{r}(1 - \frac{|n \text{ recuperados antes que } r|}{\min(R, N)})$$

donde *R* es el número de objetos relevantes, *N* el de no relevantes, *r* es un objeto relevante recuperado, y *n* es uno de los *R* primeros objetos no relevantes recuperados. *bpref* puede verse como la inversa de la proporción de objetos no relevantes que se recuperan antes que los relevantes. *bpref* y MAP estarán altamente correladas cuando se utilizan sobre juicios de relevancia completos, esto es, cuando todos los objetos han sido juzgados como relevantes o como no relevantes para cada consulta.

La medida GMAP (*geometric mean average precision*) está pensada para situaciones en las que se quiere resaltar la mejora en consultas de bajo rendimiento. Calcula la media geométrica de las precisiones medias de cada consulta, a diferencia de MAP que calcula la media aritmética. Por ejemplo, si para la consulta A se mejora de 0.02 a 0.04 y la B empeora de 0.4 a 0.38, MAP seguirá teniendo el mismo valor (0.21), mientras que GMAP mostrará algo de mejora (de 0.0447 a 0.0616), en base la siguiente fórmula que calcula la media geométrica de las precisiones medias (AP) sobre el número total de consultas (Q):

$$GMAP = \frac{\sqrt[Q]{\sum_{q=1}^{Q} AP(q)}}{Q}$$

Las principales medidas de evaluación que se utilizarán en los experimentos desarrollados en esta tesis estarán relacionadas con la precisión, tanto global (MAP), como a niveles bajos o

early precisions (P@5, P@10, P@20), ya que son las medidas utilizadas en *ImageCLEF*, foro dentro del cual se evalúa una gran parte del trabajo de esta investigación.

4.1.2 Significancia estadística

Un resultado se considera estadísticamente significativo cuando no es probable que haya sido debido al azar. Durante el desarrollo de esta tesis, se compararán diversos sistemas o configuraciones de recuperación multimedia para analizar cuál de ellos obtiene un mejor rendimiento en base a las medidas de evaluación alcanzadas por cada uno. Para las comparaciones se aplicará a los resultados de los experimentos correspondientes el test de significancia estadística. Para poder llevar a cabo este tipo de comprobaciones, será necesario disponer de los resultados de evaluación pormenorizados (para cada una de las consultas) para la medida que se desee comparar.

Un test de significancia estadística tiene asociado un nivel de significación que hace referencia a la verificación de una hipótesis (Thompson, 1994). Este nivel se define como la probabilidad de tomar la decisión de rechazar la hipótesis nula cuando ésta es verdadera (falso positivo). En general esta decisión se toma utilizando el valor *p* (o *p-valor*), que se define como la probabilidad de obtener un resultado al menos tan extremos como el que realmente se ha obtenido suponiendo que la hipótesis nula es cierta. Cuando el *p-valor* es inferior al nivel de significación (α) la hipótesis nula es rechazada. Cuanto menor es el *p-valor*, más significativo será el resultado. En resumen, cuanto menor sea el nivel de significación, más fuerte será la evidencia de que un resultado no se debe simplemente a una coincidencia.

Se rechazará la hipótesis nula si el riesgo de equivocación al asumir dicha hipótesis (*p-valor*) es superior a α. Por ejemplo, para un umbral o nivel de significación $\alpha = 0.05$, la representación gráfica de este supuesto sería:

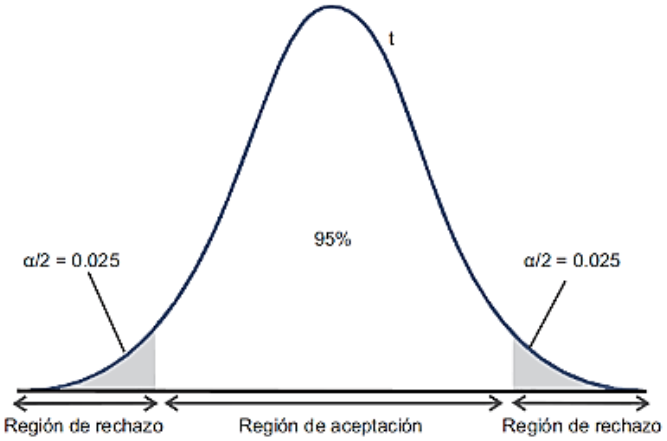

Figura 4.2. Significancia estadística

Los resultados que se analizarán en esta tesis están relacionados con la precisión en la recuperación. Se compararán distintos experimentos en base a la medida de evaluación MAP, a partir de las precisiones obtenidas en cada una de las consultas (AP). La distribución seguida por estos datos no suele ser una normal, por lo que no podrá utilizarse la prueba *t* de *Student*, que es la manera más común de llevar a cabo un análisis estadístico sobre este tipo de datos. En su lugar, se aplicará la prueba *U* de *Mann-Whitney* (también conocida como prueba de *Wilcoxon*), que es la versión no paramétrica de la prueba *t* de *Student*. Para llevar a cabo esta prueba se hace uso del entorno de programación R[15] para análisis estadístico y gráfico. Es necesario disponer de los datos a comparar (por ejemplo, los valores de MAP obtenidos para un mismo conjunto de consultas por dos sistemas distintos). A continuación, desde el entorno R, se importan los datos a comparar y se ejecuta el "Test de *Wilconxon* para muestras pareadas" disponible dentro del grupo de test estadísticos no paramétricos (el paquete *WilcoxCV* debe haber sido cargado previamente desde la biblioteca *Rcmdr*). Una vez completado el test se observa el *p-valor* obtenido para confirmar o no la significancia estadística del experimento.

4.2 Colecciones de evaluación

Las colecciones utilizadas durante el desarrollo del trabajo presentado en esta memoria, para la evaluación de las propuestas incluidas (siguiendo la metodología Cranfield) han sido:

[15] http://www.r-project.org/

- las facilitadas tras la participación del autor en el foro *ImageCLEF*, en diferentes tareas de recuperación multimedia de imágenes, y

- como prueba de concepto, la colección multimedia creada dentro del proyecto Buscamedia

4.2.1 Foro de evaluación ImageCLEF

La iniciativa *ImageCLEF* nació en el año 2003, como parte del foro de evaluación *CLEF*, en relación a la recuperación multilingüe de imágenes de diferentes dominios (Muller, Clough and Desealaers, 2010). Este foro de evaluación facilita a sus participantes el tipo de colecciones de referencia necesarias para poder llevar a cabo la investigación en el campo de la anotación y recuperación de imágenes, proponiendo un conjunto de necesidades de información multimedia (los denominados *topics*), junto con los resultados esperados (*qrels* o *groundtruth*), para que los grupos participantes desarrollen y evalúen sus sistemas. Una vez que se publican los resultados de cada grupo de investigación, se facilitan los juicios de relevancia generados para la evaluación.

El impacto académico de *ImageCLEF* ha sido muy significativo, como indica el gran número de publicaciones y citas recibidas. Las colecciones facilitadas han sido utilizadas por más de 200 grupos de investigación, facilitando la comparación de numerosas técnicas durante las diferentes ediciones, e imponiendo una sólida metodología de evaluación (Tsikrika, Herrera and Muller, 2011).

Aunque las tareas y los conjuntos de datos utilizados en las diferentes ediciones han ido cambiando con el paso de los años, los objetivos principales se han mantenido:

1) Investigar la eficacia de combinar características textuales y visuales para la tarea de recuperación de imágenes en escenarios creados para ello.

2) Recopilar y proporcionar recursos útiles como punto de referencia para sistemas de recuperación de imágenes: colecciones, *topics* y juicios de relevancia.

3) Promover el intercambio de ideas que ayuden a definir nuevas técnicas para mejorar el rendimiento de futuros sistemas de recuperación de imágenes.

Los diferentes tipos de tareas propuestas dentro del marco de *ImageCLEF*, todas ellas persiguiendo los objetivos mencionados, pueden encuadrarse en los tres grupos que se muestran a continuación:

- Recuperación *ad hoc* de imágenes. Simulan la tarea de recuperación de documentos: a partir de la descripción de las necesidades de información del usuario, se trata de encontrar todas las imágenes relevantes que sea posible (en orden de relevancia).

- Reconocimiento en imágenes de objetos y conceptos. Este tipo de tareas consisten en identificar la presencia o no de determinados objetos o conceptos (incluidos en un conjunto predefinido) en las imágenes de la colección, asignar etiquetas textuales o descripciones a las imágenes (anotación automática de imágenes), o clasificarlas en una o más clases (clasificación automática de imágenes).

- Recuperación interactiva de imágenes. En este grupo se evalúan sistemas de recuperación de imágenes utilizados por personas que interactúan con ellos. La interacción en la recuperación de imágenes puede estudiarse desde el punto de vista de cómo el sistema ayuda al usuario en la formulación de la consulta, en su traducción, o en la selección de documentos.

Las tareas organizadas dentro de estos tres grupos se han desarrollado tanto en el ámbito de las imágenes pertenecientes al dominio médico, como en el de imágenes de contenido general (archivos históricos, colecciones nuevas, imágenes de Wikipedia, etc.). El trabajo de investigación llevado a cabo en esta tesis ha sido evaluado con la participación en tareas pertenecientes a la recuperación ad hoc de imágenes, como son la tarea de recuperación de imágenes fotográficas (ediciones 2008 (Arni et al., 2009) y 2009 (Lestari, Sanderson and Clough, 2010)) y la tarea de recuperación de imágenes de Wikipedia (ediciones 2010 (Popescu, Tsikrika and Kludas, 2010) y 2011 (Tsikrika, Popescu and Kludas, 2011)).

Se describen a continuación las colecciones de imágenes utilizadas en la experimentación de las propuestas de este trabajo de tesis provenientes de la participación del autor en el citado proyecto.

4.2.1.1 Colección IAPR TC-12

Para la tarea de recuperación de imágenes fotográficas de la edición de *ImageCLEF* 2008 se hizo uso de la colección IAPR TC-12 (Grubinger et al., 2006), la cual está formada por 20.000 imágenes tomadas en diferentes lugares de todo el mundo. Contiene imágenes de diferentes deportes y actividades, fotografías de personas, animales, ciudades, paisajes, etc.

Cada una de las imágenes de la colección trae asociada información alfanumérica en un formato semiestructurado. Esta información se proporciona hasta en tres idiomas (inglés, alemán y español). La colección se encuentra disponible gratuitamente y sin restricciones de derechos de autor. Se muestra a continuación una imagen de la colección con su correspondiente anotación textual asociada, elaborada por porfesionales:

```
<DOC>
<DOCNO>annotations/00/60.eng</DOCNO>
<TITLE>Palma </TITLE>
<DESCRIPTION>two lane street with large shops on the right and smaller shops on the left; people are
walking on the sidewalk, some are crossing the street; cars are parked along the left side of the street as
well; </DESCRIPTION>
<NOTES>The main shopping street in Paraguay; </NOTES>
<LOCATION>Asunción, Paraguay </LOCATION>
<DATE>March 2002 </DATE>
<IMAGE>images/00/60.jpg </IMAGE>
<THUMBNAIL>thumbnails/00/60.jpg </THUMBNAIL>
</DOC>
```

Figura 4.3 Ejemplo de imagen de la colección IAPR TC-12

Los topics propuestos para la evaluación de los sistemas siguen el siguiente formato:

```
<top>
<num> Number: 5 </num>
<title> animals swimming </title>
<cluster>animal</cluster>
<narr> </narr>
<image> 3739.jpg </image>
<image> 4968.jpg </image>
<image> 30823.jpg </image>
</top>
```

Figura 4.4. Ejemplo de topic para colección IAPR TC-12

Cada topic de evaluación especifica en <*title*> el texto correspondiente a la necesidad de información solicitada, y en <*cluster*> el texto asociado para promover la diversidad en los

resultados a generar. Además, al tratarse de consultas multimodales, en los campos <*image*> se proporcionan imágenes de ejemplo para la recuperación de imágenes basada en contenido (CBIR).

Se enumeran a continuación las consultas propuestas. Se muestra únicamente la información textual en inglés correspondiente al campo <*title*> para cada una de las 48 consultas y el número de imágenes relevantes en la colección para cada una de ellas:

Tabla 4-1. Consultas multimedia colección IAPR TC-12

Consulta	**Parte Textual (inglés)**	**|relevantes|**
2	church with more than two towers	24
3	religious statue in the foreground	34
5	animal swimming	64
6	straight road in the USA	87
10	destinations in Venezuela	102
11	black and white photos of Russia	65
12	people observing football match	36
13	exterior view of school building	70
15	night shots of cathedrals	25
16	people in San Francisco	56
17	lighthouse at the sea	27
18	sport stadium outside Australia	45
19	exterior view of sport stadium	56
20	close-up photograph of an animal	71
21	accommodation provided by host families	69
23	sport photos from California	81
24	snowcapped building in Europe	60
28	cathedral in Ecuador	41
29	views of Sydney's world-famous landmarks	40
31	volcanoes around Quito	62
34	group picture on a beach	77
35	bird flying	87
37	sights along the Inka-Trail	95
39	people in bad weather	68
40	tourist destinations in bad weather	98
41	winter landscape in South America	135
43	sunset over water	43
44	mountains on mainland Australia	184
48	vehicle in South Korea	34
49	images of typical Australian animals	99
50	indoor photos of a church or cathedral	34
52	sports people with prizes	29
53	views of walls with unsymmetric stones	62

54	famous television (and telecommunication) towers	18
55	drawings in Peruvian deserts	81
56	photos of oxidised vehicles	28
58	seals near water	58
59	creative group pictures in Uyuni	26
60	salt heaps in salt pan	30

4.2.1.2 Colección BELGA

Esta colección, utilizada en la tarea de recuperación de imágenes fotográficas en *ImageCLEF* 2009, contiene un total de 498.920 imágenes obtenidas de la agencia de noticias BELGA, que es un motor de búsqueda de imágenes para nuevos fotógrafos.

Cada fotografía viene acompañada de su correspondiente leyenda compuesta por un texto en inglés formado por unas pocas frases. Las leyendas textuales se proporcionan sin un formato específico. Esta información textual puede contener la fecha y el lugar donde se tomó la fotografía. Se muestra a continuación un ejemplo de imagen de esta colección acompañada de la información textual que se proporciona:

Annotation: 20090126 - DENDERMONDE, BELGIUM: Lots of people pictured during a commemoration for the victims of the knife attack in Sint-Gilles, Dendermonde, Belgium, on Monday 26 January 2009. Last friday 20-Year old Kim De Gelder killed three people, one adult and two childs, in a knife attack at the children's day care center "Fabeltjesland" in Dendermonde. BELGA PHOTO BENOIT DOPPAGNE

Figura 4.5 Ejemplo de imagen y leyenda de la colección BELGA

Los *topics* propuestos están basados en el análisis de los *logs* de consulta de la agencia Belga en 2008. Gracias a esto se trata de un conjunto de consultas representativas de situaciones del

mundo real en cuanto a necesidades de información de los usuarios. Se proporcionan un total de 50 consultas de evaluación multimedia, todas ellas con la parte textual en inglés. Como particularidad, para esta edición de *ImageCLEF* 2009, cada consulta viene acompañada de la especificación de un conjunto de clases (*clusters*):

```
<top>
<num> Number: 0 </num>
<title> soccer </title>
<clusterTitle> soccer belgium </clusterTitle>
<clusterDesc> Relevant images contain photographs of the Belgium team in a soccer match. </clusterDesc>
<image> belga38/00704995.jpg </image>
<clusterTitle> spain soccer </clusterTitle>
<clusterDesc> Relevant images contain photographs of the Spain team in a soccer match. </clusterDesc>
<image> belga6/00110574.jpg </image>
<clusterTitle> beach soccer </clusterTitle>
<clusterDesc> Relevant images contain photographs of a soccer beach match. </clusterDesc>
<image> belga33/06278068.jpg </image>
<clusterTitle> italy soccer </clusterTitle>
<clusterDesc> Relevant images contain photographs of the Italy team in a soccer match. </clusterDesc>
<image> belga20/1027435.jpg </image>
<clusterTitle> soccer netherlands </clusterTitle>
<clusterDesc> Relevant images contain photographs of the Netherlands team in a soccer match or the teams in Netherlands' league.
</clusterDesc>
<image> belga10/01214810.jpg </image>
<clusterTitle> soccer -belgium -spain -beach -italy -netherlands </clusterTitle>
<clusterDesc> Relevant images contain photographs of any aspects or subtopics of soccer which are not related to the above clusters.
</clusterDesc>
<image> belga20/01404831.jpg </image>
</top>
```

Figura 4.6 Ejemplo de topic en colección BELGA

La información presente en las consultas referente a los *clusters* tiene la finalidad de ofrecer a los sistemas de recuperación la posibilidad de recuperar un conjunto de imágenes resultado que ofrezcan diversidad de contenidos, ya que ese era uno de los objetivos planteados por la tarea de recuperación de imágenes de la edición de ImageCLEF 2009.

El conjunto total de *topics o* consultas multimedia propuestas para la tarea de recuperación de imágenes fotográficas en la edición 2009 del *ImageCLEF* es el mostrado en la tabla siguiente, en la que se indica el texto de la consulta, el número de *clusters* en caso de tenerlos, y el número de imágenes relevantes para cada una de las consultas.

Tabla 4-2. Consultas multimedia colección BELGA

Q	Texto	|*clusters*|	|rel|	Q	Texto	|rel|
1	leterme	3	114	**26**	obama	174
2	fortis	4	109	**27**	anderlecht	1508
3	brussels	10	1323	**28**	mathilde	752
4	belgium	8	2210	**29**	boonen	228
5	charleroi	5	1423	**30**	china	993
6	vandeurzen	2	31	**31**	hellebaut	244
7	gevaert	3	406	**32**	nadal	125
8	koekelberg	3	119	**33**	snow	927
9	daerden	5	297	**34**	spain	1120
10	Borlee	3	35	**35**	strike	300
11	Olympic	8	1422	**36**	euro	793
12	Clinton	3	961	**37**	paris	1164
13	martens	4	181	**38**	rochus	596
14	princess	4	1045	**39**	beckham	682
15	monaco	3	246	**40**	prince	910
16	queen	4	1233	**41**	princess mathilde	672
17	tom boonen	3	198	**42**	mika	309
18	bulgaria	4	824	**43**	ellen degeneres	2
19	kim clijsters	2	1183	**44**	henin	1110
20	standard	7	1563	**45**	arsenal	721
21	princess maxima	4	192	**46**	tennis	1266
22	club brugge	5	654	**47**	ronaldo	544
23	royals	5	1202	**48**	king	618
24	paola	2	1052	**49**	madonna	197
25	mary	5	169	**50**	chelsea	740

Las primeras 25 consultas, además del texto principal para la consulta, incluyen información textual adicional para cada uno de los *clusters* a los que hacen referencia (tal y como se muestra en el ejemplo de la Figura 4.6). Además, para cada uno de los *clusters* especificados, se proporciona una imagen de ejemplo. Para las otras 25 consultas no se especifican *clusters* y se proporcionan 3 imágenes de ejemplo para cada una (parte visual de la consulta).

Los juicios de relevancia fueron realizados utilizando el sistema DIRECT[16] (*Distributed Information Retrieval Evaluation Campaign Tool*), que permite a los asesores trabajar en un

[16] http://direct.dei.unipd.it/

entorno colaborativo. 25 asesores estuvieron involucrados en la tarea. La siguiente figura muestra el número de imágenes relevantes existentes en la colección para cada consulta:

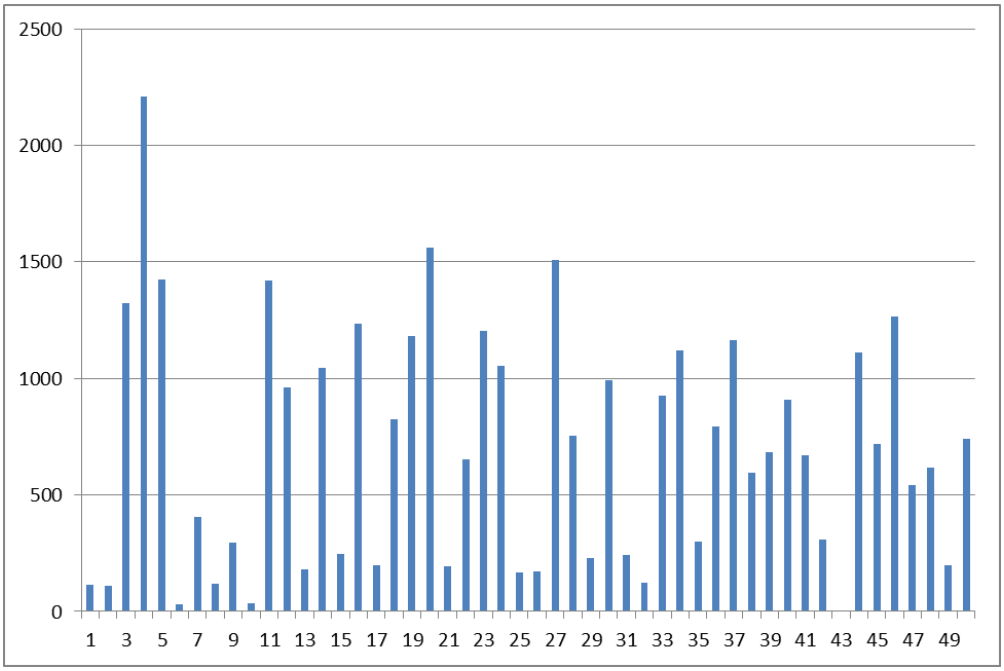

Figura 4.7. Número de imágenes relevantes por consulta (colección Belga)

En una segunda fase, y una vez que se ha identificado un conjunto de imágenes relevantes para cada consulta multimedia, los asesores se encargaron de encontrar imágenes relevantes para cada *cluster* (algunas imágenes pueden pertenecer a más de uno). La Tabla 4-2 muestra en la columna |*clusters*| el número de ellos propuestos para cada consulta.

4.2.1.3 Colección ImageCLEF 2010 Wikipedia

Esta colección ha sido utilizada durante las ediciones de *ImageCLEF* 2010 y 2011 para la tarea de recuperación de imágenes de Wikipedia. Está formada por 237.434 imágenes y sus correspondientes anotaciones textuales proporcionadas por usuarios (en inglés, francés y alemán).

Fue construida para cubrir consultas similares en inglés, francés y alemán. Con este objetivo fueron seleccionados únicamente artículos de Wikipedia con versión en cada uno de los tres idiomas, y que tuviesen al menos una imagen en cada una de las versiones. De este modo, se extrajeron 44.664 artículos del repositorio de Septiembre de 2009 de Wikipedia que contenían 265.987 imágenes. Como uno de los objetivos de la construcción de la colección era poder ser

de libre distribución, se decidió eliminar todas aquellas imágenes con derechos de autor no suficientemente claros. Tras eliminar estas imágenes, los duplicados y alguna operación de limpieza más, se obtuvo la colección final.

La distribución entre los distintos idiomas de la colección (inglés, francés y alemán) de las anotaciones textuales proporcionadas como parte de la colección de evaluación es:

Tabla 4-3. Distribución por idiomas de las anotaciones textuales

IDIOMA	Nº de imágenes
Solo ingles	70.127
Solo alemán	50.291
Solo francés	28.461
Inglés y alemán	26.880
Inglés y francés	20.747
Alemán y francés	9.646
Inglés, alemán y francés	22.899
Indeterminado	8.144
Sin anotación textual	239
Total	237.434

Se muestra a continuación un ejemplo de una imagen de la colección junto con sus correspondientes anotaciones textuales:

Figura 4.8 Ejemplo de imagen de la colección de Wikipedia

Como puede observarse en la Figura 4.8, para cada imagen se proporciona la siguiente información textual (en formato XML):

<*image*>: identificador único de la imagen (*id*), y el enlace al fichero que contiene la imagen (*file*).

<*name*>: nombre original del fichero que contiene la imagen, tal y como se encuentra en el repositorio de Wikimedia Commons, esto es, sin ningún tipo de procesamiento aplicado al texto.

<*text*>: anotaciones textuales para los casos en los que se ha identificado el idioma de las mismas (inglés, francés o alemán). Se distinguen tres campos:

<*description*>: descripción de la imagen proporcionada por los usuarios. Extraído de la página de *Wikimedia Commons* correspondiente a la imagen en el caso de que el idioma del texto haya sido explícitamente anotado de una manera normalizada.

<*comment*>: comentarios disponibles acerca de la imagen, extraídos en los mismos casos que <*description*>. Contendrá una sub-cadena del comentario general ("en crudo") mostrado más adelante.

<*caption*>: se trata del texto que acompaña a la imagen en los artículos de Wikipedia (pies de foto). También se proporciona la ruta donde se almacena el artículo en el que aparece la imagen (*article*), y su correspondiente contenido. Cuando los artículos de Wikipedia contengan imágenes sin pies de foto, este campo aparecerá vacío. Igualmente, puede darse el caso de que una imagen aparezca en más de un artículo; entonces se proporcionarán todos los pies de foto y los enlaces correspondientes.

<*comment*>: comentario general "en crudo" tal y como se encuentra en la página correspondiente a la imagen en Wikimedia Commons. Al texto no se le ha aplicado procesamiento alguno, y puede estar en uno o varios idiomas (no necesariamente inglés, francés o alemán).

<*license*>: proporciona información acerca de la licencia y los derechos de la imagen.

La información textual proporcionada en este colección tiene un alto nivel de heterogeneidad, longitudes bastante dispares, y no está exenta de ruido. Adicionalmente, la colección facilita el texto completo de los artículos de Wikipedia en los que aparecen las imágenes.

En cuanto a las consultas multimodales de evaluación (o topics), los organizadores de la tarea proporcionan 70 topics para la edición de 2010 y 50 para la de 2011. Para la selección de estos topics se analizan *logs* de consultas de búsqueda, tanto de herramientas comerciales como de los organizadores y algunos participantes. Se muestra a continuación un ejemplo de topic de la edición de 2010:

<topic>

<number> 68 </number>

<title xml:lang="en"> historic castle <title>
<title xml:lang="de"> historisches schloss <title>
<title xml:lang="fr"> château fort historique <title>

<image> 3691767116_caa1648fee.jpg </image>
<image> 4155315506_545e3dc590.jpg </image>

<narrative> We like to find pictures of historic castles.
The castle should be of the robust, well-fortified kind.
Palaces and chateaus are not relevant.
</narrative>

</topic>

Figura 4.9 Ejemplo de topic para la colección de Wikipedia 2010

Como puede observarse la consulta es multimodal/multimedia, con una parte textual para cada uno de los tres idiomas de la colección, y una parte visual con dos imágenes de ejemplo (obtenidas a partir de Flickr, bajo licencia *Creative Commons*). El campo <*title*> está formado por los términos textuales que utilizaría un usuario real para realizar una consulta. Una vez que un usuario se diese cuenta de que la consulta textual no es suficiente, podría decidir añadir imágenes de ejemplo para completar una consulta multimedia (serían las imágenes referenciadas en los campos <*image*>). El campo <*narrative*> debe contener una clara y precisa descripción de la necesidad de información, con el objetivo de determinar sin ambigüedad cuándo una determinada imagen satisface o no una necesidad de información. Esta información será utilizada posteriormente en el proceso de generación de los juicios de relevancia. La relevancia será binaria, esto es, una imagen puede ser juzgada como relevante o

como no relevante para cada una de los topics de la colección. Este juicio es manual y llevado a cabo por parte de evaluadores, que se encargarán cada uno de una consulta en particular.

El conjunto total de consultas proporcionadas en la tarea de recuperación de imágenes de Wikipedia de *ImageCLEF* 2010 se muestra en la siguiente tabla, donde se indica la parte textual (en inglés) de cada consulta, así como el número de imágenes ejemplos proporcionados en cada caso (|ejemplos|), y el número de imágenes relevantes en la colección (|relevantes|):

Tabla 4-4. Consultas multimedia *ImageCLEF* 2010

Consulta	**Parte Textual (inglés)**	**|ejemplos|**	**|relevantes|**
1	fractals	2	317
2	cockpit of an airplane	1	87
3	basketball game close up	2	116
4	Christmas tree	2	22
5	Oktoberfest beer tent	2	9
6	solar panels	2	101
7	lightning in the sky	1	43
8	tennis player on court	2	393
9	flying hot air balloon	2	30
10	horseman	2	96
11	landline telephone	1	27
12	DNA helix	1	39
13	trains and locomotives	2	687
14	videogames screenshot	2	114
15	cyclist	2	176
16	spider with cobweb	2	27
17	beach volleyball	2	7
18	stars and galaxies	2	384
19	lochs in Scotland	1	53
20	mountains with sky	2	969
21	Chernobyl disaster ruins	2	17
22	sharks underwater	2	27
23	emoticon smiley	2	8
24	Rorschach black and white	1	6
25	Shiva painting or sculpture	2	29
26	brain scan	2	24
27	active volcano with ash cloud	1	75
28	palm trees	2	71
29	desert scenery	2	247
30	harbour	2	454
31	yellow buses	1	50

32	people laughing	2	51
33	close up of antenna	2	90
34	people playing guitar	2	348
35	race car	2	852
36	portrait of Jintao Hu	1	5
37	close up of bottles	1	237
38	baseball game	1	140
39	cactus in desert	1	13
40	ferrari red	1	485
41	polar bear	2	46
42	Paintings related to cubism	2	23
43	skyscraper in daylight	2	362
44	saturn	2	81
45	snowy winter landscape	2	376
46	sailboat	1	181
47	soccer stadium	1	366
48	civil airplane	1	633
49	surfing on waves	1	38
50	portraits of people	2	1727
51	aerial pictures of landscapes	3	678
52	satellite image	2	875
53	ISS international space station	1	178
54	launching space shuttle	1	102
55	building site	1	125
56	musician on stage	1	568
57	road street signs	2	305
58	red fruits	2	146
59	cities at night	3	528
60	notes on music sheet	1	233
61	earth from space	2	89
62	Shopping in a market	2	224
63	postage stamp	3	866
64	woman in red dress	2	57
65	sea sunset or sunrise	1	116
66	bridges in daylight	2	793
67	white house with garden	1	77
68	historic castle	2	605
69	red tomato	1	33
70	close up of trees	2	603

Los juicios de relevancia correspondientes a las consultas de evaluación propuestas, son construidos siguiendo una aproximación basada en un *pooling* al estilo TREC de profundidad 100. La relevancia es binaria, esto es, cada imagen de la colección en relevante o no lo es para cada consulta. La evaluación se lleva a cabo por tres de los grupos participantes y por los

organizadores de la tarea durante un periodo de cuatro semanas posterior al envío de experimentos con los resultados por parte de cada uno de los grupos participantes.

Para la edición de la tarea de recuperación de imágenes de Wikipedia correspondiente al *ImageCLEF* 2011, los organizadores proponen un nuevo conjunto de consultas multimedia de evaluación. Las nuevas consultas se muestran en la siguiente tabla:

Tabla 4-5. Consultas multimedia *ImageCLEF* 2011

Consulta	**Parte Textual (inglés)**	**|ejemplos|**	**|relevantes|**
71	colored Volkswagen beetles	5	50
72	skeleton of dinosaur	5	116
73	graffiti street art on walls	5	95
74	white ballet dress	5	49
75	flock of sheep	5	34
76	playing cards	5	47
77	cola bottles or cans	5	24
78	kissing couple	5	33
79	heart shaped	5	34
80	wolf close up	4	25
81	golf player on green	5	22
82	model train scenery	5	40
83	red or black mini cooper	5	10
84	Sagrada Familia in Barcelona	5	7
85	Beijing bird nest	5	12
86	KISS live	5	11
87	boxing match	5	45
88	portrait of Segolene Royal	5	10
89	Elvis Presley	4	7
90	gondola in Venice	5	62
91	freestyle jumps with bmx or motor bike	5	18
92	air race	5	12
93	cable car	5	47
94	roller coaster wide shot	5	155
95	photo of real butterflies	5	112
96	shake hands	5	77
97	round cakes	5	43
98	illustrations of Alice’s adventures in Wonderland	4	21
99	drawings of skeletons	5	95
100	brown bear	5	46
101	fountain with jet of water in daylight	5	141
102	black cat	5	20
103	dragon relief or sculpture	5	41
104	portrait of Che Guevara	4	13

105	chinese characters	5	316
106	family tree	5	76
107	sunflower close up	5	13
108	carnival in Rio	5	37
109	snowshoe hiking	5	12
110	male color portrait	5	596
111	two euro coins	5	58
112	yellow flames	5	92
113	map of Europe	5	267
114	diver underwater	5	33
115	flying bird	5	115
116	houses in mountains	5	105
117	red roses	4	27
118	flag of UK	4	12
119	satellite image of desert	4	93
120	bar codes	4	14

Se incluye a continuación la Tabla 4-6 con un resumen de las características principales de los dos conjuntos de consultas proporcionadas en las 2 ediciones de la tarea de recuperación de imágenes de Wikipedia del *ImageCLEF* en las que se utilizó la colección de evaluación aquí descrita:

Tabla 4-6. Resumen consultas *ImageCLEF*

	2010	**2011**
Número de consultas	70	50
Número de imágenes por consulta	1,68	4,84
Número de términos por consulta	2,7	3,08
Número de imágenes relevantes por consulta	252,3	68,8
Número de entidades nombradas por consulta	0,1	0,26

La principal diferencia se encuentra en el número de imágenes de ejemplo proporcionadas para cada una de las consultas multimedia propuestas. Puede observarse que en la edición 2010 este número es algo bajo (1,68). No se dispone ni siquiera de dos ejemplos visuales para cada consulta, lo que dificulta el trabajo de los algoritmos de recuperación basados en las características visuales de bajo nivel (CBIR). El conjunto para 2011 está formado por consultas multimedia con una media de 4 o 5 imágenes de ejemplo por consulta (4,84). Esta ampliación fue sugerida por el grupo de investigación del que forma parte el autor de esta tesis, y aceptada positivamente por los organizadores de la tarea. Otras distinciones

remarcables son el número de imágenes relevantes existentes en la colección para cada consulta, bastante mayor en la edición 2010, y la cantidad de entidades nombradas identificadas en la parte textual de las consultas multimedia, que es algo mayor en la edición de 2011.

4.2.2 Escenario Buscamedia

El proyecto de investigación Buscamedia[17], enmarcado dentro del programa CENIT[18], está enfocado al avance en las áreas de semántica, producción audiovisual y distribución de media enriquecida con independencia de redes y terminales de consumo y con el objetivo de crear un buscador semántico multimedia único. Buscamedia pretende el desarrollo de tecnologías de búsqueda multimedia y gestión automatizada de software que permita crear una base sólida para el desarrollo de una amplia gama de servicios de muy diversa índole en el ecosistema audiovisual.

Dentro del proyecto Buscamedia se aborda la creación de un corpus multimedia para la experimentación y evaluación de los distintos activos que se desarrollan a lo largo del proyecto. Es el denominado corpus Deportes20, cuya composición puede verse en la Tabla 4-7.

[17] http://www.cenitbuscamedia.es/
[18] http://www.cdti.es/

Tabla 4-7. Composición corpus Deportes20

	Información disponible	N	Idioma
Videos - Eventos deportivos	• Vídeos (11vídeos) • Keyframes • Metadatos • Objetos (10videos) • Transcripciones (11 videos, ASR castellano, sin indexar)	21	ca
Videos - Informativos deportivos	• Vídeos • Subtítulos + Texto sobreimpreso • Transcripciones • Logos + Moscas (4 videos)	10	es
Noticias deportivas relacionadas	• Noticias de la colección textual (título, descripción, sección, fecha, titulo_noticia, entradilla, cuerpo, foto)	62	ca (30) es (30) eu (2)
Páginas web con noticias relacionadas	• Páginas web con noticias relacionadas (título, descripción, keywords, sección, titulo_noticia, entradilla, cuerpo, fecha, autor, lugar)	34	es (30) en (1) ca (3)
Videos deportivos de Youtube	• Vídeos • Keyframes • Metadatos	4	es (3) ca (1)

La colección Deportes20 está compuesta por 131 recursos multimedia: 21 vídeos de eventos deportivos y 10 vídeos de informativos. También contiene 62 noticias deportivas procedentes de un corpus textual propio del proyecto (21.632 elementos), y de 34 páginas web externas seleccionadas por su relación con el contenido de los vídeos del corpus. Durante la evolución del proyecto se añaden otros 4 nuevos vídeos sobre temas relacionados con el deporte. Fue desarrollado en el año 2011.

Para poder utilizar esta colección como prueba de concepto de propuestas válidas en ImageCLEF, se desarrollan un conjunto de consultas de evaluación, junto con sus correspondientes juicios de relevancia. La siguiente tabla muestra el conjunto de consultas:

Tabla 4-8. Conjunto de consultas (corpus Deportes20)

Consultas para corpus Deportes20			
1	Hat trick de Fernando Torres	**18**	vídeos de jugadores en un campo de balonmano
2	Fernando Alonso en el GP de Corea	**19**	partidos en el Santiago Bernabéu
3	Iker Casillas	**20**	duelo entre Dani Pedrosa y Jorge Lorenzo
4	Equipo patrocinado por bwin	**21**	premios Príncipe de Asturias a la selección de futbol
5	zeppelín en evento deportivo	**22**	cuatro goles de Cristiano Ronaldo
6	cristiano ronaldo en murcia	**23**	balón de oro
7	quién es el líder de la liga (noticias)	**24**	viaje del Barcelona a Ceuta
8	quién es el líder del mundial (vídeos)	**25**	Barcelona balonmano
9	competicions de ping pong	**26**	noticias sobre el nastic / nastic de tarragona
11	Casademont Girona	**27**	alberto contador tour
12	noticias y vídeos de Mourinho	**28**	declaraciones de Pep Guardiola
13	goles messi	**29**	Gobierno de Aragón
14	ciclistes participants en el mundial de fons en carretera	**30**	gol de Villa a Wembley
15	enfrentamientos entre Juande Ramos y Ernesto Valverde	**31**	delanteros argentinos
16	resultados juegos olímpicos de Pekín	**32**	Cristiano Ronaldo y Mourinho
17	partidos en los que aparece Ronaldinho	**33**	Real Madrid - CSKA

Se construyen manualmente los juicios de relevancia y se almacenan en formato TREC, indicando para cada consulta aquellos documentos multimedia que satisfarían las necesidades de información del usuario. De este modo, y generando ficheros de resultados en el mismo formato, podrán evaluarse automáticamente los resultados obtenidos haciendo uso de la herramienta *trec_eval* (descrita en el apartado 4.3.1).

Tabla 4-9. Fichero con juicios de relevancia para Buscamedia (corpus Deportes20)

1 0 1084 1 *2 0 isid_LaSexta24-10(1)-1 1* *2 0 LaSexta24-10(1)_MpegCorteNativo 1* *2 0 TVE24-10(2)_MpegCorteNativo 1* *2 0 es_-476971287733223.8966 1* *2 0 es_-220971287614491.5767 1* *2 0 es_412391287586141.6275 1* *2 0 ca_-442291288090269.0691 1* *3 0 1066 1* *3 0 ca_-105161288260231.0311 1* *3 0 1384 1* *3 0 es_942181287412673.577 1* *3 0 es_-165191287473623.7848 1* *3 0 es_-750331287412269.3459 1* *4 0 TVE24-10(1)_MpegCorteNativo 1* *4 0 LaSexta24-10(2)_MpegCorteNativo 1* *4 0 LaSexta24-10(1)_MpegCorteNativo 1* *5 0 1246 1* *6 0 TVE26-10(1)_MpegCorteNativo 1* *6 0 isisd_TVE24-10(1)-2 1* *6 0 isid_LaSexta24-10(2)-2 1* *6 0 isisd_TVE24-10(1)-1 1* *6 0 TMAD25-10_MpegCorteNativo 1* *6 0 isid_tmad25-10-1 1* *6 0 isid_LaSexta24-10(2)-1 1* *6 0 isid_TVE126-10-1 1* *7 0 es_311531287440744.2095 1* *7 0 isid_TVE126-10-1 1* *7 0 es_-31851287698436.4843 1* *8 0 LaSexta24-10(1)_MpegCorteNativo 1* *9 0 1686 1* *11 0 1246 1* *12 0 TVE24-10(1)_MpegCorteNativo 1* *12 0 isid_tmad25-10-1 1* *12 0 es_942181287412673.577 1* *12 0 es_-165191287473623.7848 1* *12 0 es_-750331287412269.3459 1* *13 0 es_-348111287642174.9278 1* *13 0 es_-181571287614666.3425 1* *13 0 es_-194031287699328.48 1* *13 0 es_-385351287642554.1676 1* *13 0 TVE24-10(1)_MpegCorteNativo 1* *14 0 1071 1* *15 0 1070 1* *16 0 2774 1* *16 0 2359 1* *16 0 ccma_2774-2 1* *16 0 ccma_2539-2 1* *16 0 ccma_2539-1 1* *16 0 ccma_2774-1 1* *16 0 2554 1* *16 0 2545 1* *16 0 2427 1* *17 0 1388 1* *17 0 1693 1* *17 0 ccma_1388-1 1* *17 0 ccma_1693-1 1* *17 0 ccma_1693-2 1*	*17 0 ccma_1388-2 1* *17 0 es_-278421287556479.2761 1* *18 0 1688 1* *19 0 1384 1* *19 0 es_-278421287556479.2761 1* *19 0 es_308041287614601.8487 1* *19 0 isisd_TVE24-10(1)-1 1* *19 0 ccma_1384-1 1* *19 0 ccma_1384-2 1* *19 0 isisd_TVE24-10(1)-2 1* *19 0 isid_LaSexta24-10(2)-1 1* *19 0 TVE24-10(1)_MpegCorteNativo 1* *20 0 1102 1* *21 0 T521-10_MpegCorteNativo 1* *21 0 es_942181287412673.577 1* *21 0 isid_TVE121-10-1 1* *21 0 es_-165191287473623.7848 1* *21 0 es_-750331287412269.3459 1* *21 0 TVE21-10_MpegCorteNativo 1* *22 0 LaSexta24-10(2)_MpegCorteNativo 1* *22 0 LaSexta24-10(1)_MpegCorteNativo 1* *22 0 TVE24-10(1)_MpegCorteNativo 1* *23 0 TVE26-10(2)_MpegCorteNativo 1* *23 0 es_-151641287654810.0007 1* *23 0 isid_LaSexta25-10-2 1* *23 0 TVE24-10(1)_MpegCorteNativo 1* *23 0 ca_-105161288260231.0311 1* *24 0 TVE26-10(1)_MpegCorteNativo 1* *24 0 LaSexta24-10(2)_MpegCorteNativo 1* *24 0 ca_483651288109759.1378 1* *24 0 ca_74701288090266.3447 1* *24 0 isid_TVE126-10-1 1* *24 0 Lasexta25-10_MpegCorteNativo 1* *25 0 ccma_1688-2 1* *25 0 ccma_1388-2 1* *25 0 ccma_1688-1 1* *25 0 1688 1* *26 0 ccma_1383-1 1* *26 0 ccma_1383-2 1* *26 0 ca_-355891287208808.0787 1* *27 0 es_146491287558324.9146 1* *27 0 es_-377571287558882.4456 1* *27 0 2539 1* *27 0 es_119351287412173.895 1* *27 0 ca_-20501287557452.4111 1* *28 0 es_204981287559351.7146 1* *28 0 es_942181287412673.577 1* *28 0 es_311531287440744.2095 1* *29 0 TVE24-10(1)_MpegCorteNativo 1* *30 0 video4-Villa 1* *31 0 es_-348111287642174.9278 1* *31 0 video3–Messi 1* *31 0 isid_TVE121-10-2 1* *32 0 TVE24-10(1)_MpegCorteNativo 1* *32 0 es_-165191287473623.7848 1* *32 0 video2-CristianoRonaldo 1* *33 0 video1-Mourinho 1*

El proceso de anotación llevado a cabo sobre este corpus es:

- Anotación de imágenes
    - Reconocimiento de caras

  - Reconocimiento de logos y moscas
  - Reconocimiento de objetos en general
  - Anotación de texto en *keyframes* de vídeo
- Anotación de textos
  - Reconocimiento de textos en vídeo: subtítulos y texto sobreimpreso
  - Texto libre
- Anotación de audio (habla)
  - Transcripción textual, que será tratada como texto normal
  - Anotación semántica de recursos de audio hablados (conceptos)

El componente encargado de manejar el proceso de anotación tomará el archivo de entrada (URL donde se encuentra físicamente el recurso) y en base a su media (modalidad de vídeo, texto, audio o imagen) ejecutará un procesado u otro. La siguiente figura muestra, a modo de ejemplo, el proceso seguido para el caso de la anotación multimedia de las imágenes (o *keyframes* de los vídeos). Será anotado en función del tipo de recurso multimedia a tratar.

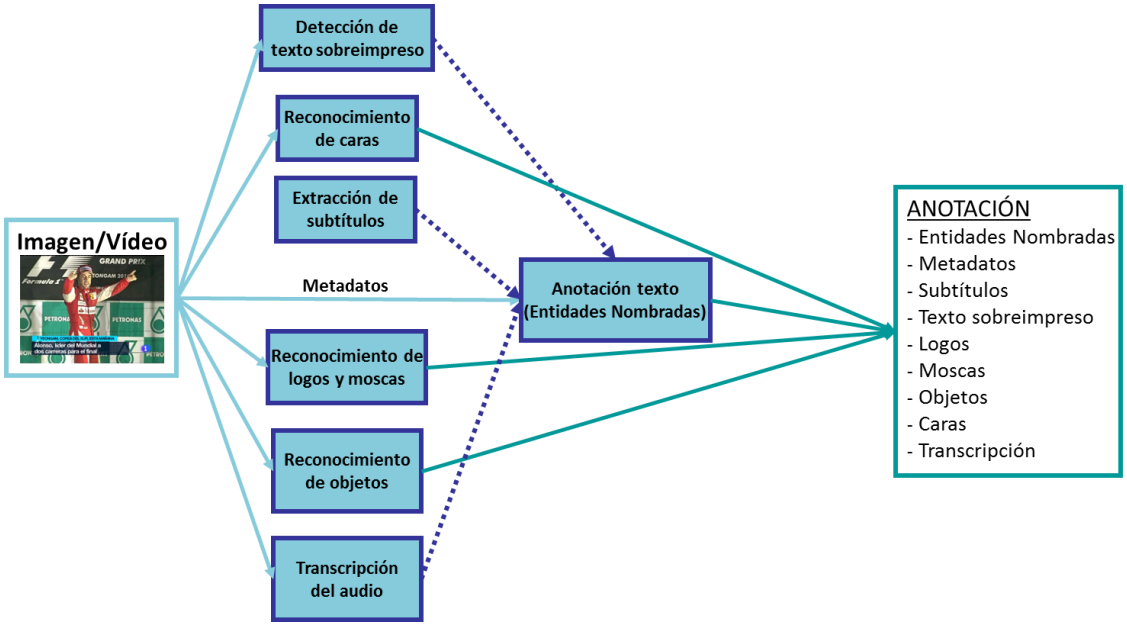

Figura 4.10. Anotación de imágenes/vídeos en Buscamedia

La anotación de los recursos multimedia de tipo imagen o vídeo generan un fichero XML de salida que contiene en forma de texto diferentes elementos detectados en el recurso (logos, moscas, objetos, etc.). También se incluyen las entidades nombradas identificadas en los

metadatos asociados al recurso. Toda esta información, junto con los metadatos textuales originales será indexada para su posterior recuperación.

La siguiente figura muestra la interfaz correspondiente a los componentes de anotación y fusión temprana en Buscamedia. A partir de identificador del recurso multimedia, el componente *Anotador* se encargará de anotar la información multimedia obtenida en cada caso. El componente de fusión combinará las anotaciones completadas construyendo un único fichero XML en el que se dispondrá de la anotación multimedia final del recurso identificado.

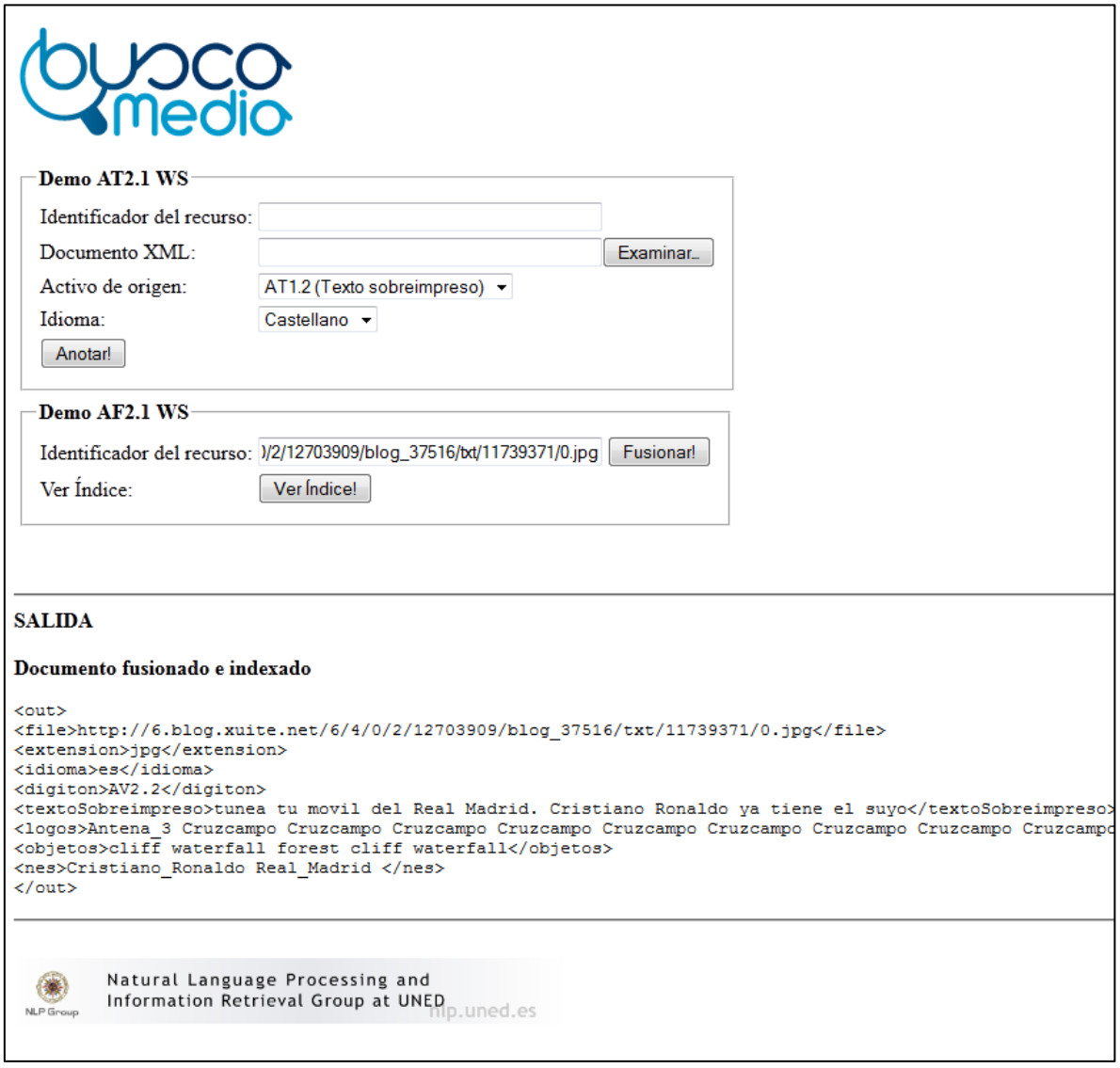

Figura 4.11. Fusión de anotaciones multimedia en Buscamedia

A continuación se muestra, a modo de ejemplo, un fichero XML con la anotación correspondiente a uno de los documentos que forman parte del corpus Deportes20 (un vídeo de un informativo deportivo).

```
<?xml version="1.0" encoding="UTF-8"?>
<out>
   <file>TVE24-10(2)</file>
   <extension>mpeg</extension>
   <idioma>es</idioma>
   <digiton>ISID</digiton>
   <textoSobreimpreso>YEONGAM, COREA DEL SUR, ESTA MAÑANA. Alonso, líder del Mundial
a dos carrearas para el final</textoSobreimpreso>
   <logos>Petronas, Formula 1, F1</logos>
   <objetos>persona</objetos>
   <transcripcion>Fernando Alonso a colocar al líder del Mundial y todo gracias a... </ transcripcion >
   <audio_conceptos>formula1</audio_conceptos>
   <subtitulos>formula1</subtitulos>
   <nes>Corea del Sur, Fernando Alonso, Mundial de Corea</nes>
</out>
```

Figura 4.12. Ejemplo de salida tras anotación multimedia en Buscamedia

Esta salida será indexada por la herramienta Lucene, siguiendo un esquema basado en campos. Para la recuperación se utiliza la función de ranking BM25F, que extiende a BM25 para documentos estructurados, es decir, formados por campos. Los factores de empuje utilizados son iguales para todos los campos, con lo que la diferencia efectiva entre la aportación de cada campo se deberá a las diferencias encontradas en los valores medios de las longitudes de los campos. La modificación en cuanto al peso de cada campo introducida por la función de ranking BM25F es la encargada de ponderar la importancia de cada una de las fuentes o modos de información multimedia en el proceso de fusión temprana de anotaciones.

4.3 Herramientas utilizadas

Se dedica este apartado a describir brevemente las herramientas utilizadas durante el desarrollo de esta tesis.

4.3.1 IDRA (*InDexing and Retrieving Automatically*)

La herramienta IDRA (Granados Muñoz, García-Serrano and Goñi Menoyo, 2009), desarrollada e implementada como parte del desarrollo de esta tesis, es utilizada para el preprocesamiento textual de las anotaciones asociadas a las imágenes, la indexación/recuperación de imágenes basada en la información textual, y la fusión multimedia entre resultados textuales y visuales en base a algoritmos de fusión tardía (*late fusion*).

Una completa descripción de esta herramienta puede encontrarse en el Anexo de esta tesis, donde se describen todas las funcionalidades ofrecidas y algunas capturas de la interfaz gráfica de la herramienta.

Figura 4.13. Herramienta IDRA

La herramienta IDRA se encuentra disponible como software libre en *SourceForge* bajo licencia GPL, y accesible desde https://sourceforge.net/projects/idraproject/. Hasta el momento se han registrado más de 1.000 descargas de la herramienta.

4.3.2 Lucene

La herramienta Apache Lucene[19] (Gospodnetic and Hatcher, 2004) es un API (*Application Programming Interface*) de código abierto para la recuperación de información, apoyada por la *Apache Software Foundation*[20] que se distribuye por *Apache Software License*.

[19] http://lucene.apache.org/
[20] http://www.apache.org/

Inicialmente fue desarrollada en Java pero actualmente pueden encontrarse distintas versiones en otros lenguajes de programación como Perl (llamada Plucene), Ruby (Ferret y RubyLucene), Python (Pylucene), C++ (CLucene), C (Lucene4c) o .NET (NLucene y Lucene.Net).

La herramienta permite la indexación y búsqueda de documentos así como la revisión de ortografía, marcado de palabras claves y capacidades de análisis y *tokenización* avanzadas. Lucene provee de las tecnologías necesarias para el procesamiento de documentos en inglés pero también permite indexar y buscar sobre otras lenguas. Es importante tener en cuenta que existen distintos tipos de analizadores ya desarrollados en idiomas como chino, alemán, francés o ruso. Para el español existe un analizador *snowball* que maneja cadenas de texto, implementando algoritmos de *stemming*.

Para comprender el funcionamiento de Lucene es importante entender los conceptos fundamentales del sistema: índice (*index*), documento (*document*), campo (*field*), término (*term*), y segmento (*segment*). Básicamente, un índice contiene un conjunto de documentos y un documento es una secuencia de campos. Por su parte, un campo es una secuencia de términos y un término es un par de cadenas de caracteres que representan el nombre del campo y el valor o el texto dentro del campo. Los índices generados por la herramienta pueden estar compuestos por múltiples subíndices o segmentos. Cada uno de estos segmentos es un índice independiente sobre el que se puede buscar de forma separada. Esto permite que existan dos mecanismos básicos de desarrollo de índices: bien creando nuevos segmentos para nuevos documentos o bien, fusionando segmentos existentes.

Por una parte, Lucene permite indexar cualquier dato que pueda ser convertido a texto, lo cual implica que puede realizar la indexación de páginas web, mensajes de correo electrónico, documentos Microsoft Word, ficheros PDF, HTML o cualquier documento del que pueda extraerse información textual. Además, permite la creación de índices invertidos que muestran para un determinado término los documentos que lo contienen. Los índices almacenan estadísticas sobre los términos para hacer la búsqueda basada en términos más eficiente.

En relación a las búsquedas que permite Lucene, éste se basa en los modelos vectorial y booleano puro e incluye la posibilidad de incluir entre sus consultas operadores booleanos, búsquedas basadas en campos, búsquedas comodines, etc. Es importante señalar que Lucene

presenta algunas carencias para implementar un sistema completo de recuperación de información. Por ejemplo, no incluye rastreadores o recopiladores de información ("*crawling*") o analizadores HTML. Sin embargo, existen varios proyectos paralelos que trabajan para desarrollar estas funcionalidades.

4.3.3 Herramienta de evaluación *trec_eval*

Es la herramienta estándar utilizada por la comunidad TREC para evaluar los experimentos de recuperación *ad hoc*. Los resultados referentes a las medidas de evaluación de los experimentos que componen esta tesis están calculados haciendo uso de la herramienta *trec_eval*.

Para poder ejecutar este programa de evaluación es necesario disponer de un fichero con los juicios de relevancia (*qrels* o *ground truth*) para las consultas asociadas a la colección de evaluación, y de otro fichero con los resultados de recuperación generados por el experimento a evaluar. Ambos ficheros deberán respetar un formato predefinido.

El fichero de juicios de relevancia tendrá una línea por cada juicio emitido. Cada línea estará compuesta por varios campos (separados por espacios en blanco o tabuladores) que indicarán el identificador de la consulta para el que se emite el juicio, el número de iteración (que para las tareas de recuperación automática es siempre 0), el identificador del documento u objeto multimedia juzgado, y el valor binario del juicio (1 si se considera relevante, 0 en otro caso). Un ejemplo de fichero con juicios de relevancia para un conjunto de consultas (con identificadores 71, 72,...) se muestra en la siguiente figura:

```
71 0 106844 0
71 0 106866 1
71 0 106867 1
71 0 106885 0
71 0 106886 0
71 0 106921 1
72 0 66377 0
72 0 70663 1
72 0 70693 0
72 0 70694 0
72 0 70695 1
...
```

Figura 4.14. Ejemplo de fichero de juicios de relevancia

Los ficheros de resultados a evaluar también deberán respetar el formato definido para *trec_eval*. En este caso cada línea se corresponderá con un documento u objeto multimedia (por ejemplo, una imagen) recuperado para una determinada consulta. Cada línea contendrá información sobre: 1) identificador de la consulta, 2) número de iteración (siempre a 1), 3) identificador del documento u objeto multimedia resultado, 4) posición o *ranking* (desde 0), 5) valor de relevancia o similitud (*score*), y 6) identificador del experimento. En la siguiente figura se muestra un ejemplo de fichero de resultados:

```
71  1  106974  0  0.3854749996855928   MAXmerge-norm
71  1  106970  1  0.36142492617993766  MAXmerge-norm
71  1  106869  2  0.33741365567582965  MAXmerge-norm
71  1  14951   3  0.33383112865101594  MAXmerge-norm
...
72  1  75230   0  0.5004706258494835   MAXmerge-norm
72  1  171280  1  0.4954405974978174   MAXmerge-norm
72  1  70695   2  0.47829883264033396  MAXmerge-norm
...
```

Figura 4.15. Ejemplo de fichero de resultados (formato TREC)

Con estos dos ficheros correctamente formados, se podrá lanzar el programa *trec_eval* de la siguiente manera (se indican los argumentos más utilizados, para ver todos usar *-h* en la herramienta):

trec_eval [-q] [-c] [-M <num>] [-] fichero_juicios fichero_resultados

donde:

- q: para obtener los valores de evaluación para cada una de las consultas independientemente, aparte de los globales.
- c: para evaluar en base a todas las consultas existentes en el fichero de juicios de relevancia, y no solo en la intersección de estos con los resultados.
- M <num>: para indicar el número máximo de documentos por consulta a tener en cuenta en la evaluación.

4.3.4 Plataforma gráfica para visualización de resultados

Esta plataforma fue desarrollada como parte de un proyecto fin de carrera (Martinez and Benavent, 2010), en la Escuela Técnica Superior de la Universidad de Valencia. En dicha plataforma se encuentran alojadas varias colecciones (alrededor de 16) de las utilizadas en las diferentes tareas y ediciones de *ImageCLEF*. La plataforma ofrece diversas funcionalidades, entre las que cabe destacar las siguientes:

- Visualización del contenido de las colecciones, esto es, de las imágenes. Además, pinchando sobre cada una de ellas podrá verse su información textual asociada.

- Visualización de las consultas multimedia, donde pueden observarse las distintas imágenes de ejemplo proporcionadas para cada consulta, junto con la información textual.

- Visualización de resultados de experimentos. Podrá verse el conjunto de imágenes recuperadas para cada consulta según el experimento analizado.

- Visualización de los juicios de relevancia.

La plataforma está disponible en http://imageclef.uv.es/ImageClef/index.php (accesible mediante identificación).

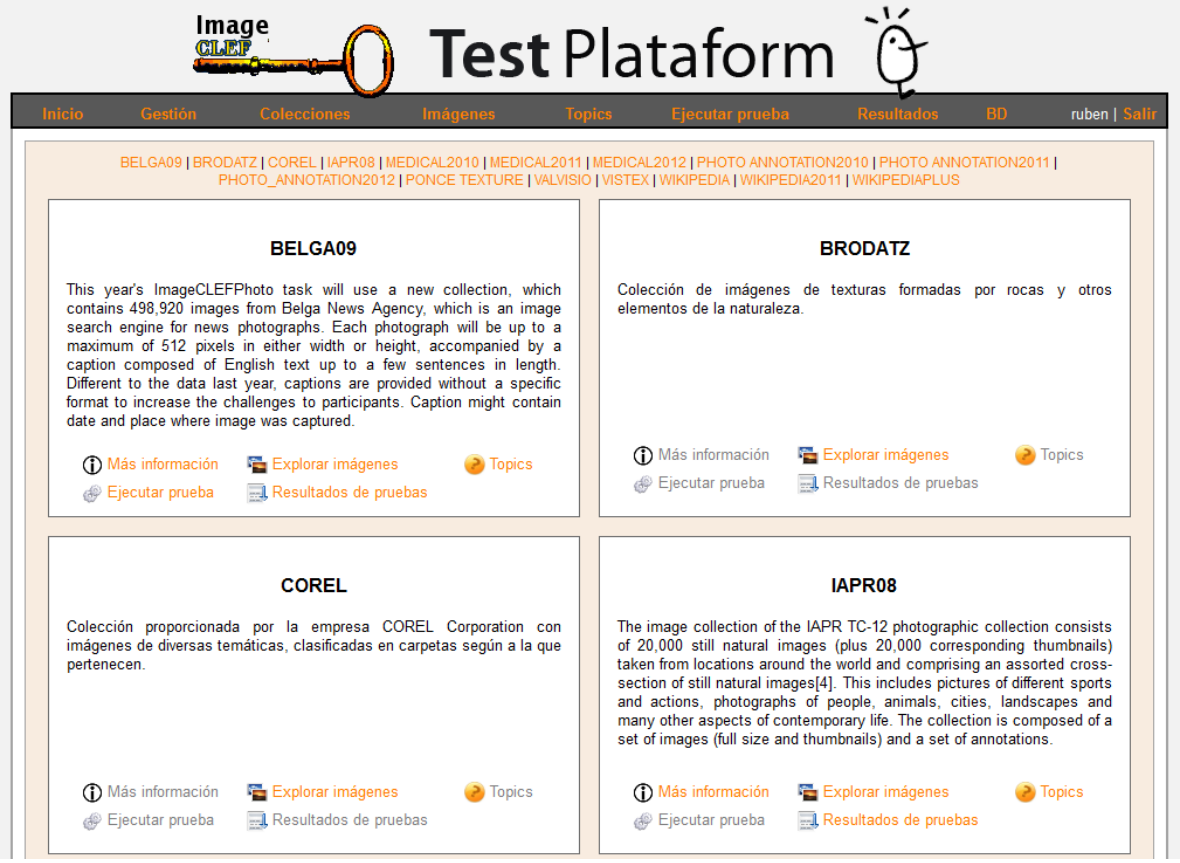

Figura 4.16. Interfaz de la plataforma de visualización

Esta plataforma ha resultado especialmente útil durante el desarrollo del trabajo de esta tesis, por ejemplo para analizar tanto las imágenes que forman parte de cada base de datos como las diferentes consultas multimedia asociadas a cada una de ellas. También ha resultado muy útil a la hora de compartir distintos tipos de experimentos y poder evaluar y analizar los resultados obtenidos por cada uno de ellos desde cualquier lugar y de "un vistazo".

4.3.5 Activos Buscamedia

Dentro del marco de trabajo del proyecto Buscamedia se desarrollan diferentes funcionalidades o componentes software relacionados con la búsqueda y recuperación de elementos multimedia pertenecientes al corpus Deportes20 (descrito en la sección 4.2.2) o a la colección de imágenes de Wikipedia (sección 4.2.1). Se describen a continuación dos de los componentes implementados como parte del proyecto, que son utilizados para la prueba de concepto de algunas de las contribuciones de esta memoria

El primero de los componentes, denominado *Buscador Configurable*, permite al usuario lanzar consultas seleccionado manualmente los campos de búsqueda a utilizar, así como otros parámetros como el idioma o el operador lógico a aplicar. Los campos de búsqueda están relacionados con las distintas fuentes de información procedentes de cada modalidad en la colección Deportes20 (por ejemplo: metadatos o entidades nombradas para el texto, subtítulos u objetos para el vídeo, o transcripciones para el audio). La siguiente figura muestra la interfaz diseñada para el uso de este tipo de búsqueda:

Figura 4.17. Interfaz Buscamedia: *Búsqueda Configurable*

Por ejemplo, para la consulta textual "*Cristiano Ronaldo*" y seleccionando todos los campos de búsqueda disponibles, así como los diferentes tipos de recursos a recuperar, la interfaz mostraría los resultados como se muestra en la siguiente figura. Cuando se pincha sobre uno de los resultados obtenidos, el recurso correspondiente es abierto (en este caso, una parte de un vídeo a partir de un punto en el que se menciona a Cristiano Ronaldo).

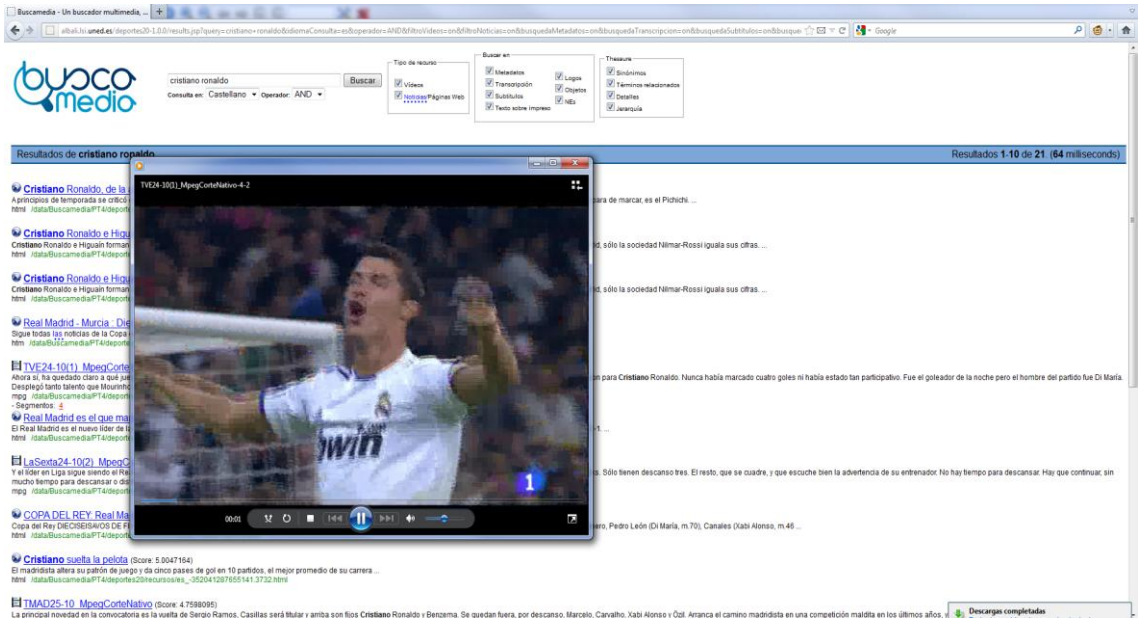

Figura 4.18. Resultados Interfaz *Búsqueda Configurable*

También se desarrolló un interfaz de búsqueda para la colección de imágenes de Wikipedia, que es finalmente la colección de evaluación utilizada para la principal contribución de esta tesis. Se muestra a continuación una captura de dicha interfaz:

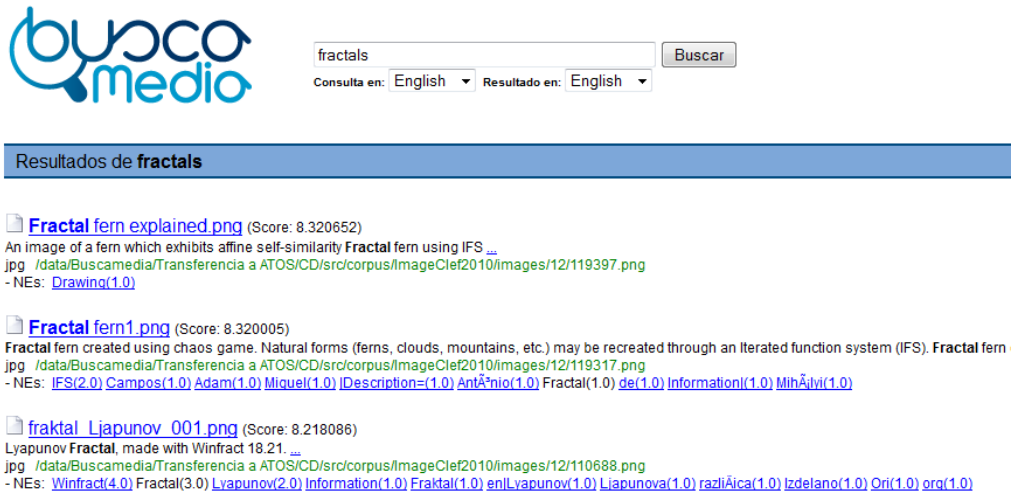

Figura 4.19. Interfaz Buscamedia para *ImageCLEF*

El otro componente corresponde al tipo de búsqueda denominado *Búsqueda Automática*, basado en un manejador que se encarga de analizar el texto introducido por el usuario en la consulta y, a partir de él, tomar las decisiones que considere oportunas en base a un conjunto

de reglas predefinidas. La interfaz implementada para que el usuario introduzca el texto de su consulta y, opcionalmente, el idioma de la misma es:

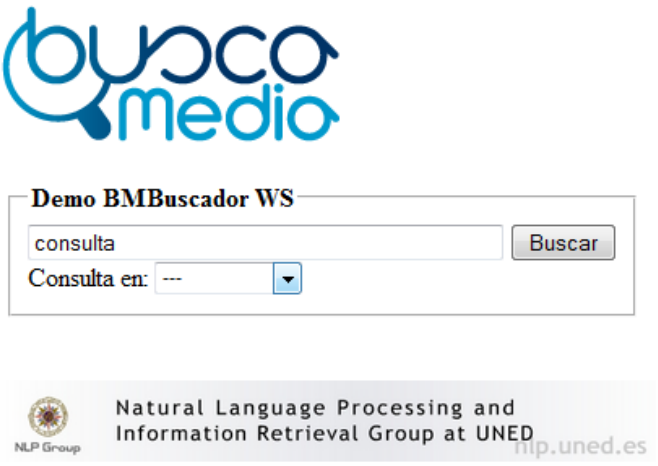

Figura 4.20. Interfaz Buscamedia: *Búsqueda Automática*

Las reglas de decisión que guían el comportamiento de la búsqueda automática se basan en el análisis del texto de la consulta introducida por el usuario. Este análisis consiste en identificar el idioma de la consulta, las entidades nombradas, objetos y logos detectables por los componentes software correspondientes. También se intentará reconocer automáticamente el tipo de recurso deseado por el usuario. El manejador de la *Búsqueda Automática* deberá seleccionar los campos del índice sobre los que lanzar las consultas, así como la ponderación o empuje (*boost*) para cada uno de ellos. Para las consultas multitérmino también deberá decidir qué operador lógico aplicar.

Se utiliza el detector de idioma *Stilus Lang*[21] proporcionado por la empresa Daedalus[22]. Hay que mencionar la dificultad de este tipo de herramientas para identificar correctamente el idioma de un texto cuando este está formado por pocas palabras, que en muchas ocasiones tampoco respetan la corrección gramatical.

El reconocimiento de las entidades nombradas se realiza con la herramienta *Stilus NER*[23] para los idiomas inglés y castellano, y con la herramienta *FreeLing*[24] para el catalán. En caso de

[21] http://www.daedalus.es/productos/stilus/stilus-lang/
[22] http://www.daedalus.es/
[23] http://www.daedalus.es/productos/stilus/stilus-ner/

que el idioma no haya sido especificado por el usuario ni identificado por el manejador, se utilizará el detector de Stilus para castellano, al ser el idioma predominante en la colección utilizada.

A partir del análisis de la consulta textual, se configuran los parámetros de búsqueda de la consulta final con las siguientes reglas:

- Idioma de la consulta. Si es conocido (indicado por el usuario o detectado con *Stilus Lang*), se realizará el preprocesamiento lingüístico correspondiente para ese idioma. En caso de ser desconocido, se aplicará el preprocesamiento correspondiente a los cuatro idiomas del corpus (castellano, catalán, euskera e inglés) y se concatenarán los resultados obtenidos.

- Operador lógico de búsqueda. Si se trata de una consulta multipalabra siempre se utilizará el operador OR, ya que recuperará también los resultados que se obtendrían con el operador AND y añadiría otros adicionales que no cumpliesen con todos los términos de la consulta.

- Entidades Nombradas. Si en la consulta se reconocen entidades, se lanzará una consulta con el campo del índice correspondiente a las entidades ponderado al doble. De este modo, se pretende dar mayor peso a la información textual referente a las entidades. Los campos de información textual general, como los metadatos, las trascripciones, los subtítulos, o el texto sobre impreso, también serán utilizados para la búsqueda

- Objetos. En caso de identificar en la consulta textual alguna referencia a los objetos de una lista (detectables por componentes visuales), se incluirá el campo del índice <*objetos*> entre los seleccionados para lanzar la consulta final. También se dará mayor importancia a dicho campo que a los demás (con una ponderación doble) para aprovechar la información multimedia disponible (esta parte de la consulta se proporciona a partir de la imagen). Los campos de información textual general también serán utilizados para la búsqueda.

---

[24] http://nlp.lsi.upc.edu/freeling/

- Logos. En el caso de detectar algún logo en la consulta introducida, el tratamiento será el mismo que en el caso anterior en el que se detectaba un objeto, esto es, se pondera al doble el campo correspondiente para aprovechar la información disponible.

- Los campos correspondientes a la considerada como información textual general de los documentos multimedia serán siempre utilizados en la configuración de todas las búsquedas lanzadas.

- Tipo de recurso. Se pretende identificar en el texto de la consulta introducida por el usuario el tipo de recurso multimedia que este prefiere recuperar como resultado. En el caso de detectar términos como "vídeo/s", "noticia/s" o "web/s", la búsqueda quedará restringida a recuperar documentos multimedia del tipo correspondiente. Con este fin se utiliza el campo del índice en el que se almacena la extensión de los diferentes recursos multimedia: "htm", "html" y "txt" para recursos de tipo noticia o página web, y extensiones "wmv", "mpg", "mpeg" o "xml" para recursos de tipo vídeo.

Este conjunto de reglas serán las que guíen el funcionamiento de la funcionalidad de *Búsqueda Automá*tica, tras el proceso previo de análisis de la consulta textual.

PARTE 2: PROPUESTA Y EXPERIMENTACIÓN

En esta segunda parte se presenta una detallada descripción de la propuesta realizada, junto con el conjunto de experimentos llevados a cabo. Inicialmente se explica la idea y motivación de la aproximación propuesta, describiendo los distintos componentes o fases que son necesarios. A continuación se muestra el sistema de recuperación multimedia diseñado e implementado, que se utilizará para ejecutar los experimentos. Finalmente, y tras mostrar los resultados obtenidos en los distintos experimentos desarrollados, se incluyen las conclusiones obtenidas que estarán relacionadas con las aportaciones de este trabajo.

Capítulo 5 Propuesta: Fusión Multimedia Semántica Tardía

La propuesta de Fusión Multimedia Semántica Tardía (*Late Semantic Multimedia Fusion*, LSMF) (Benavent et al., 2013) está enfocada a aprovechar de la mejor manera posible la información disponible en una colección multimedia cuando se aborda una tarea de recuperación.

5.1 Justificación

En el caso de las imágenes, los tipos de información presentes en las colecciones son la información visual (las imágenes propiamente dichas), y la información textual (anotaciones asociadas a las imágenes u otra información adjunta). El reto es afrontar el problema del *semantic gap* que, como se describe en el apartado 2.2, se refiere a la distancia que hay entre

la forma de entender la información percibida por el usuario a partir de una imagen y la percibida por el sistema a partir de las características de bajo nivel de la misma imagen (histogramas de color, niveles de textura, etc.). Esta brecha semántica puede observarse claramente cuando se realizan consultas del tipo *Query by Example* (guiada por una o varias imágenes), incluso cuando se utiliza el motor de búsqueda de imágenes de Google[25] como se muestra en la siguiente figura, en la que se utiliza una de las imágenes proporcionadas en *ImageCLEF 2011* como parte de una de las consultas, concretamente la correspondiente a "*diver underwater*" mostrada en la Figura **5.8**:

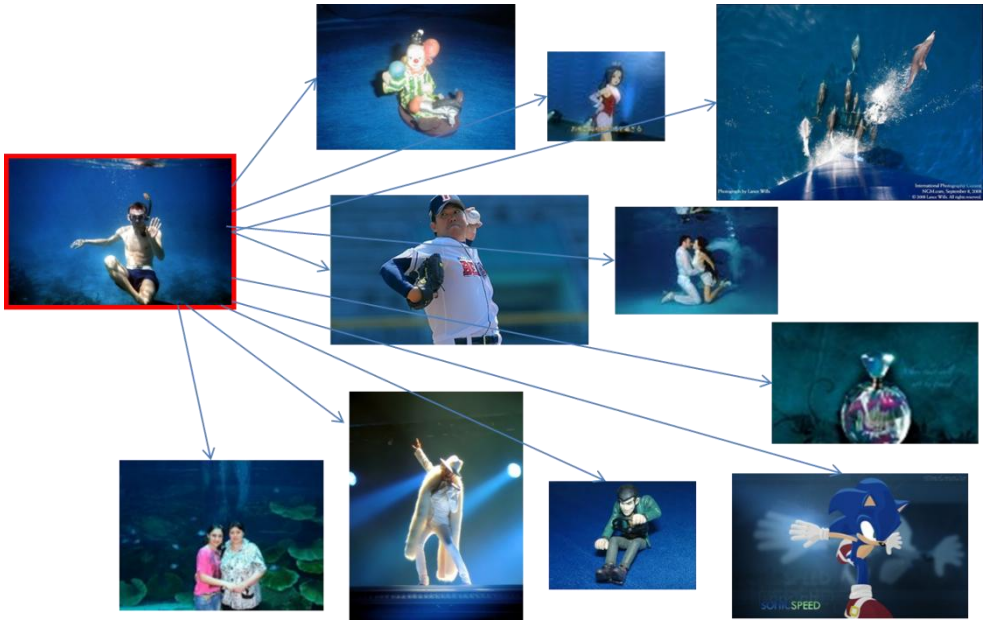

Figura 5.1. Resultados de búsqueda con *Google Images*

Puede verse como, aunque las características visuales de bajo nivel de las imágenes recuperadas sean parecidas a las del ejemplo (color, textura, etc.), ninguna de las imágenes está relacionada semánticamente con lo que el usuario pretende encontrar, esto es, a alguien buceando. Queda clara, por tanto, la necesidad de avanzar en la recuperación multimedia de imágenes basada en la colaboración entre los modos textual y visual de las mismas. Como se verá más adelante (apartado 5.3), el problema del *semantic gap* también puede darse cuando se lleva a cabo una recuperación puramente textual.

[25] http://images.google.com/

El objetivo propuesto trata de aumentar el rendimiento de los sistemas de recuperación de imágenes basada en texto (TBIR) que, como se ha visto en el estado del arte, es la aproximación que mejor ha funcionado hasta el momento, incluyendo información o resultados obtenidos por sistemas de recuperación de imágenes basadas en el contenido visual (CBIR). En los últimos cinco años el autor de esta tesis ha investigado en esta área concreta, y las soluciones propuestas en la literatura para combinar la información procedente de ambas modalidades (visual y textual) se basan en la fusión multimedia, bien a nivel de características (*early fusion*) o a nivel de decisiones (*late fusion*). Es la segunda alternativa la que obtiene mejores resultados. La siguiente figura muestra el esquema habitual de fusión multimedia tardía aplicado a la recuperación de imágenes, mediante la combinación de listas de resultados:

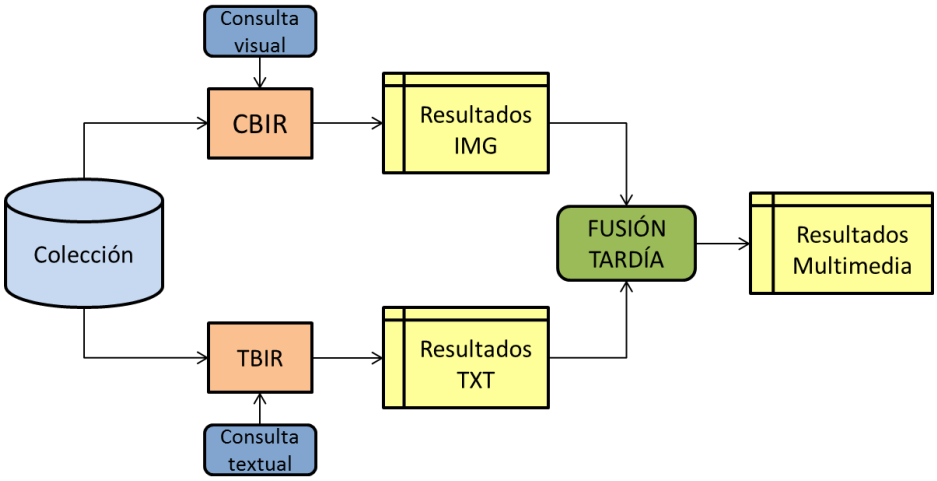

Figura 5.2. Esquema habitual de la Fusión Multimedia Tardía

La principal aportación de esta tesis es la propuesta de una estrategia de fusión multimedia asimétrica (en contraposición a las técnicas de fusión existentes) basada en un prefiltrado textual, planteada por primera vez en (Granados et al., 2011) dentro del marco de trabajo del foro de evaluación *ImageCLEF 2011*.

Esta aportación se basa en la inclusión de una fase inicial de filtrado de la colección original sobre la base de la información textual, previa a la recuperación basada en el contenido visual y, por consiguiente, a la fase de fusión, dando como resultado una lista de resultados de CBIR sobre la colección filtrada textualmente. Posteriormente, se llevará a cabo la fusión multimedia a nivel de decisiones (*late fusion*) entre las listas de resultados obtenidas desde

cada modalidad: TBIR y CBIR. Los resultados desde el sistema CBIR se obtendrán trabajando únicamente sobre la colección prefiltrada, con las consiguientes ventajas en relación al tiempo de procesamiento y cálculo (aspectos críticos para CBIR). Esta etapa inicial se diferencia de la técnica de *image reranking*, vista en el estado del arte, en que en esta propuesta se realiza una fusión posterior, incrementando de esta forma la influencia del análisis TBIR frente al CBIR, como se corresponde con la constatación de que la información textual es más descriptiva y discriminante en general que la visual (color, textura, etc.). Además, la fase de prefiltrado proporcionará al sistema CBIR de un conjunto de imágenes de contraejemplo, un segundo aspecto crítico, no tenido en cuenta en muchas aproximaciones, pero que mejora los resultados visuales.

Esta aproximación no había sido reconocida explícitamente (Csurka, Clinchant and Popescu, 2011) hasta que en la edición de *ImageCLEF 2011* fue presentada individual y simultáneamente por dos grupos: Xerox y UNED-UV (Granados et al., 2011), obteniendo los dos mejores resultados en la tarea de recuperación de imágenes de Wikipedia (Tsikrika, Popescu and Kludas, 2011). Ya en ediciones anteriores (2008, 2009 y 2010) el grupo UNED-UV había presentado en el mismo foro de evaluación estrategias similares que se encaminaban a la aproximación final de LSMF (*Late Semantic Multimedia Fusion*), conceptualmente similar a la propuesta por el grupo Xerox en (Clinchant, Csurka and Ah-Pine, 2011), donde se referencia ya el trabajo relacionado y que forma parte de esta tesis (Benavent et al., 2010). La siguiente figura muestra un esquema de la aproximación propuesta:

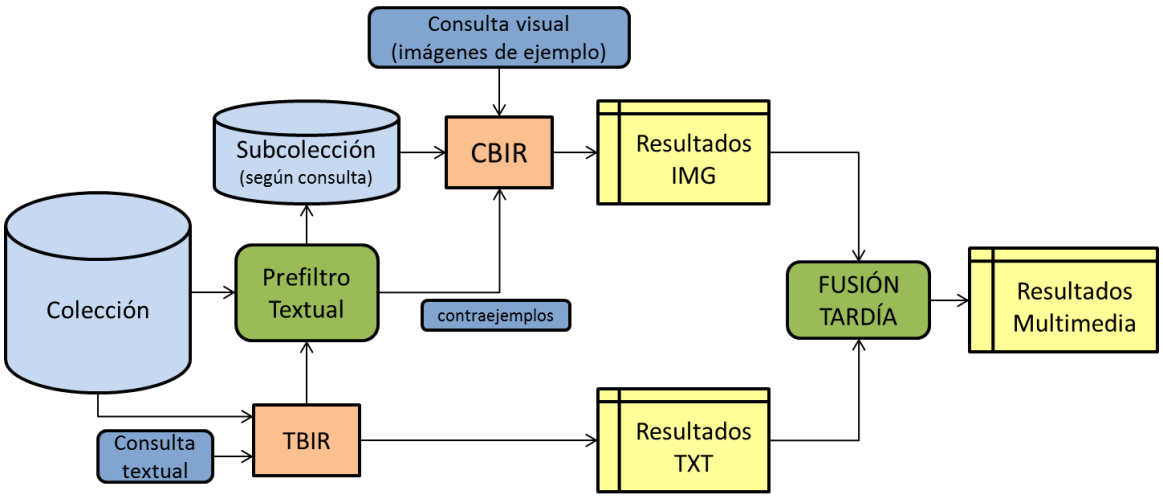

Figura 5.3. Fusión Multimedia Semántica Tardía (*Late Semantic Multimedia Fusion*, LSMF)

Esta técnica surge del hecho de que la información disponible en cada uno de los modos (texto e imagen) está expresada con muy distinto nivel semántico (y también de representación), por lo que parece lógico que la combinación entre ambas no deba llevarse a cabo de una manera equilibrada o simétrica, como hacen muchas de las técnicas de fusión tardía (Figura 5.1). Lo que se propone es aplicar una estrategia de fusión asimétrica o balanceada, como puede verse en la Figura 5.3, basada en la certeza de que el sistema TBIR inicialmente captura mejor la información semántica de las imágenes y de las consultas, ya que la información textual se encuentra en un nivel semántico más alto que la información visual, como se muestra en la Figura 5.4 (las características textuales suelen ser términos o conceptos que, por lo general, tienen una carga semántica mayor que la que pueda tener un histograma de color o cualquier otra característica visual de una imagen). Esta afirmación se prueba analizando los resultados de evaluación obtenidos en base a los sistemas monomodales de recuperación (TBIR y CBIR), como se verá en la sección dedicada a la experimentación.

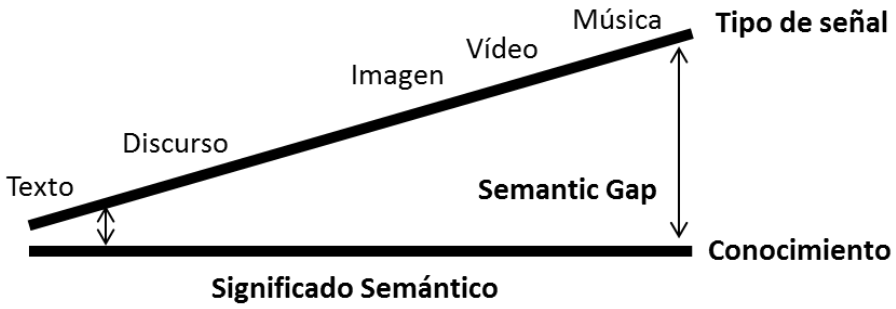

Figura 5.4. Escala semántica según tipo de información multimedia **(Baeza-Yates and Ribeiro-Neto, 2011)**

En la siguiente figura puede apreciarse la diferencia fundamental entre la aproximación clásica de fusión multimedia tardía.

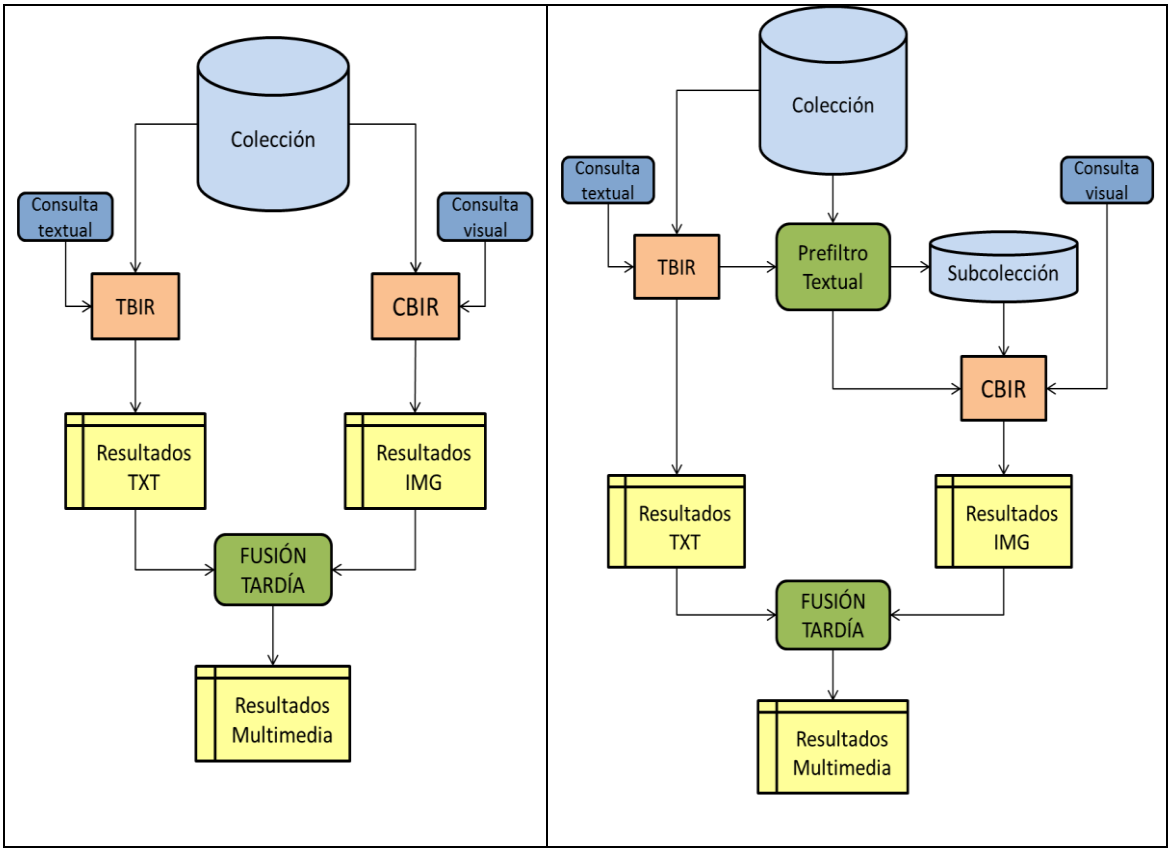

Figura 5.5. Comparación Fusión Tardía con LSMF

A continuación se detalla la contribución mencionada a la recuperación multimedia, describiendo la justificación, funcionamiento y ventajas de la fase de prefiltrado en el apartado 5.2, la evaluación de distintos algoritmos de fusión multimedia tardía

(independientemente, o como parte de la estrategia LSMF) en el apartado 5.3, y el funcionamiento global de la técnica LSMF en el apartado 5.4.

5.2 Prefiltro textual

El análisis del rendimiento de los sistemas TBIR en comparación con los CBIR cuando se afronta una tarea de recuperación multimedia de imágenes, deja claro que la información textual captura de mejor manera el contenido semántico de las mismas, en relación a los descriptores visuales utilizados por CBIR. En el Capítulo 7 de esta memoria, dedicado a la experimentación, pueden verse diferentes casos en los que se observa este hecho. Como se ha dicho, este comportamiento se asocia principalmente a la gran diferencia entre la información semántica aportada (Figura 5.4) por las anotaciones textuales con respecto a las características visuales de bajo nivel, produciendo el problema del *semantic gap*.

La propuesta de incluir una fase inicial de prefiltrado sobre la base de la información textual y previa a la actuación del sistema CBIR, facilitará la tarea de recuperación al sistema CBIR en varios sentidos:

1) el sistema CBIR trabajará únicamente sobre un subconjunto de imágenes que estarán, teóricamente, más relacionadas semánticamente con las consultas del usuario.

2) la reducción de la colección original simplificará al sistema CBIR el costoso proceso de extracción de características visuales y cálculo de semejanzas de las imágenes, lo que hará escalable la tarea de recuperación sobre colecciones grandes.

3) además proporcionará un subconjunto de imágenes no relacionadas semánticamente con la consulta, que el sistema CBIR podrá utilizar como ejemplos negativos (contraejemplos) para sus algoritmos de recuperación.

4) y todo ello sin una pérdida de cobertura alta (solo quedarán descartadas imágenes similares visualmente a la consulta, pero que no disponen de anotación textual, o no son de calidad o relevantes textualmente para la consulta), como se prueba en los resultados.

La Figura 5.6 muestra un adelanto de los resultados que se analizarán en la fase de experimentación, los cuales apoyan los comentarios anteriores. En ella puede observarse la

mejora de los resultados del sistema CBIR cuando se incluye la fase de prefiltrado textual (MAP: 0.14% → 6.18%). Se muestra además la considerable reducción de la colección de imágenes (casi hasta el 2% de su tamaño original), sin que esto suponga una perdida demasiado alta de imágenes relevantes (la colección prefiltrada mantiene una cobertura de casi el 83 %).

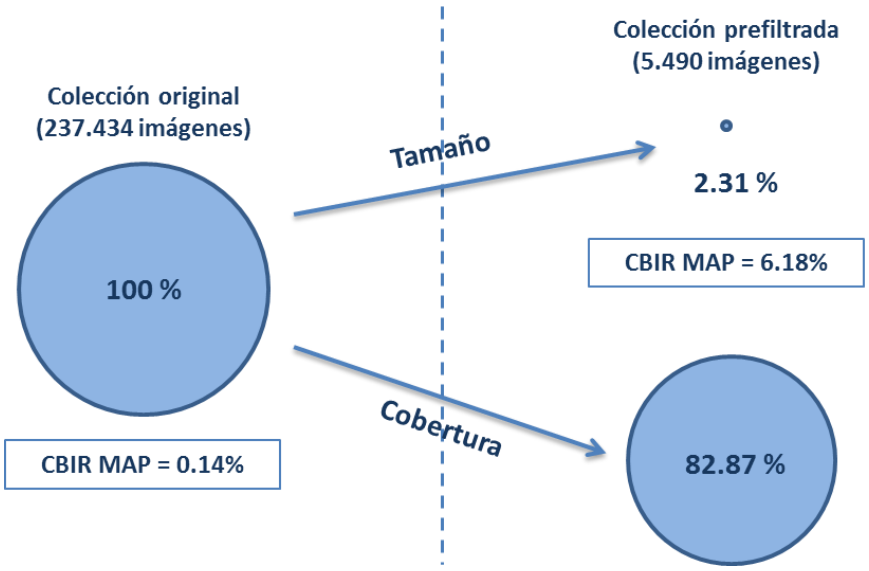

Figura 5.6. Beneficios del uso del prefiltro textual (*ImageCLEF* 2011)

A continuación se muestra un ejemplo gráfico en el que se puede observar la mejora introducida por el uso del prefiltro textual en cuanto a los resultados visuales (CBIR). Para ello se sigue con la consulta multimedia "*diver underwater*" ("buceador bajo el agua"), utilizada anteriormente y cuya parte visual se muestra más adelante (Figura 5.8). La Figura 5.7 muestra las primeras 15 imágenes recuperadas para el caso de la aproximación visual pura CBIR en comparación con cuando se añade la etapa de prefiltrado en base a la información textual:

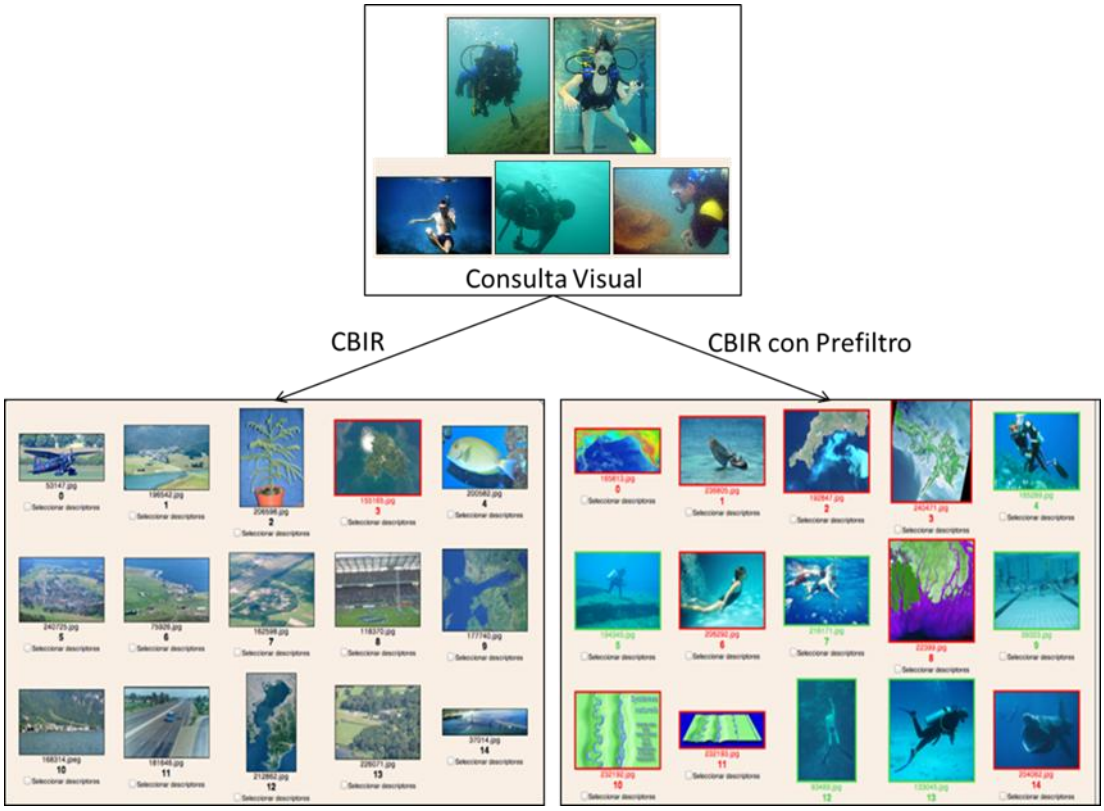

Figura 5.7. Mejora CBIR con Prefiltro

Pueden apreciarse los beneficios de incluir el prefiltro textual ya que, mientras los resultados de CBIR puro no obtienen ninguna imagen relevante (aunque todas son parecidas visualmente en términos de color, textura, forma, etc.), con la fase previa de prefiltrado se consigue recuperar entre las primeras 15 imágenes un total de 7 en las que aparecen personas buceando. Esta mejora se explica gracias a que el sistema textual ha eliminado de la colección de imágenes original aquellas que, según sus características textuales, poco o nada tienen que ver con la necesidad de información expresada en la parte textual de la consulta multimedia. Es así como al proceso global de recuperación visual se le añade un mayor nivel de semántica, proporcionado por el sistema de recuperación basado en texto.

Analizando los resultados de evaluación para la consulta de ejemplo mostrada, se obtiene un $MAP = 0.35\%$ cuando el sistema CBIR trabaja sobre la colección original de imágenes, y un $MAP = 21.42\%$ cuando lo hace sobre la colección prefiltrada. Por otro lado, $P@100$ en el primer caso es igual a 0 (no se recupera ninguna imagen relevante entre los 100 primeros resultados), mientras que si se introduce la fase de prefiltrado se obtiene $P@100 = 18\%$. Los valores de precisión baja pueden observarse directamente en la Figura 5.7: $P@10$ sin prefiltro

sería 0, y con prefiltro alcanzaría un 40%. También los valores de cobertura son mejorados, pasando de un 30.30% a un 72.72%. Estos datos dejan clara la mejoría obtenida en los resultados visuales (CBIR) cuando se incluye la etapa de prefiltrado en base a la información textual de la consulta, sin olvidar la simplificación del proceso de recuperación visual gracias a la reducción de la colección de imágenes: para el caso concreto de esta consulta se pasa de un total de 237.434 imágenes a trabajar únicamente con 1.665 (0.7%). Todo este análisis se verá refrendado en el Capítulo 7 de experimentación, donde se evaluará el comportamiento del prefiltro trabajando sobre la colección de evaluación del *ImageCLEFwiki* 2010 y 2011, tanto por separado como conjuntamente.

El funcionamiento del prefiltro textual se describe a continuación. Siguiendo el esquema de la Figura 5.5, recibe como entradas la colección original de imágenes y la lista de resultados generada por el sistema TBIR. El prefiltro construirá una subcolección de imágenes formada solo por aquellas que tengan alguna semejanza con la parte textual de las consultas. De este modo se obtendrá una colección de imágenes restringida por la consulta, en la que todas las imágenes tendrán una relación semántica con las necesidades de información expresadas por el usuario. Esta subcolección será sobre la que trabajará el sistema CBIR con las técnicas de recuperación basadas en las características visuales (o de bajo nivel) de las imágenes.

La recuperación CBIR llevada a cabo tras la fase de prefiltrado difiere de la técnica de *image reranking* vista en el estado del arte. Esta técnica consiste en reordenar los resultados procedentes de la recuperación textual (TBIR) sobre la base de la relevancia o *scores* calculados por el subsistema visual (CBIR), lo que suele mejorar notablemente los resultados puramente visuales, pero no los textuales. Dentro de la propuesta aquí planteada, también se mejorarán los resultados textuales al completar la fase de prefiltrado con la fusión multimedia.

5.3 Algoritmos de Fusión Multimedia Tardía (*Late Fusion*)

Se implementan y evalúan, dentro del trabajo llevado a cabo para esta tesis, varios algoritmos de fusión tardía (*late fusion*) o a nivel de decisiones, y se realiza un profundo análisis de las ventajas e inconvenientes de utilizar cada uno de ellos. Se comparará el rendimiento de la recuperación multimedia tanto cuando estos algoritmos se integran dentro de la técnica propuesta de fusión multimedia semántica tardía (LSMF), como cuando se utilizan como parte de una aproximación clásica de fusión tardía (*late fusion*). El objetivo de la fusión tardía

será aprovechar las particularidades y la complementariedad existente entre los distintos modos para mejorar los resultados textuales que, como se ha visto en el estado del arte, son los que mejor rendimiento ofrecen.

Un caso ilustrativo de cómo las decisiones tomadas por el sistema CBIR pueden ayudar a las del TBIR puede verse en el siguiente ejemplo (Figura 5.8), correspondiente a la tarea de recuperación de imágenes de Wikipedia de *ImageCLEF* (edición 2011), donde se muestra la parte visual de la consulta multimedia (la parte textual es "*Diver underwater*", siguiendo con el mismo ejemplo del apartado anterior).

Figura 5.8. Parte visual (ejemplos) de la consulta "*Diver underwater*"

Se ha observado que en este caso el sistema TBIR recupera en sus primeras posiciones una imagen relacionada textualmente con la consulta, pero que en realidad no se considera relevante para las necesidades de información del usuario. Esta imagen (Figura 5.9) muestra una insignia que es entregada a los buceadores en determinadas ocasiones, por lo que en lo que respecta a sus anotaciones textuales sí está relacionada con el buceo o el submarinismo. Se trata de otro ejemplo de *semantic gap*, en este caso de la recuperación textual, ya que las anotaciones de la imagen hablan de buceo (insignia) pero no concretamente de alguien buceando que es lo que busca el usuario.

Ubicació: 21/205947.jpg

NAME: Sukeltajasiivet.jpg
DESCRIPTION: Finnish Defence Forces breast insignia for divers, 1st class, also called "diver wings"
COMMENT: {{Information |Description={{fi|Suomen puolustusvoimien sukeltajan koulutusmerkki, I luokka, ns. sukeltajasiivet}} {{en|Finnish Defence Forces breast insignia for divers, 1st class, also called "diver wings"}} |Source=http://www.mil.fi/merivoimat/joukot/e)
CAPTION: Finnish diver insignia

Figura 5.9. Imagen (y su anotación textual) de la colección de Wikipedia

La complementariedad existente entre la información textual y visual presente en las imágenes, de la que tratan de sacar provecho las técnicas de fusión multimedia, puede apreciarse claramente en este ejemplo. Un buen algoritmo de fusión debería ser capaz de eliminar la imagen de la insignia de entre las imágenes recuperadas (o al menos de las primeras posiciones de la lista), ya que el sistema CBIR no encontrará ningún parecido visual entre dicha imagen y las proporcionadas como ejemplo (mostradas en la Figura 5.8).

Los algoritmos implementados y analizados en este trabajo son los descritos en el siguiente capítulo, y serán componentes de la herramienta de recuperación implementada, descrita en el apartado 6.4. Estos algoritmos de fusión tardía son: *MaxMerge*, *FilterN*, *OWA*, *Product*, y *Enrich*.

5.4 Fusión Multimedia Semántica Tardía (LSMF)

La técnica de fusión multimedia propuesta en esta tesis combina el uso de un prefiltro textual como paso previo a la recuperación visual, con la posterior utilización de algoritmos de fusión tardía entre los resultados obtenidos por los sistemas monomodales (TBIR y CBIR).

El análisis de la literatura referente a la recuperación de información multimedia, y en concreto la de imágenes, llevado a cabo en la parte del estado del arte, muestra cómo las aproximaciones empleadas están basadas en sistemas de fusión multimedia que utilizan esquemas simétricos de fusión tardía (tales como la media aritmética o *combMNZ*). Esto no se corresponde con el hecho de que la semántica de una imagen es más difícil de extraer a partir de su contenido visual que desde sus anotaciones textuales. Tampoco resulta suficiente con ponderar la influencia de cada modalidad mediante funciones de agregación con diferentes pesos. Es por esto que la propuesta presentada en este trabajo consiste en un esquema de fusión multimedia asimétrico donde, en una fase inicial, se restringirá la colección de imágenes original en base a sus características textuales. Aprovechando este prefiltrado, se enriquecerá la consulta visual con contraejemplos o "imágenes negativas", esto es, imágenes

que no están relacionadas con la consulta (aspecto interesante para los algoritmos de recuperación visual). Esto permitirá al sistema CBIR trabajar únicamente con imágenes relacionadas con las consultas (según TBIR), y aprovechar la información visual de los contraejemplos proporcionados por el prefiltro. Posteriormente, y tras la recuperación monomodal de cada una de las fuentes de información tras el prefiltrado, se aplicará un esquema de fusión tardía para aprovechar la complementariedad existente entre ambos modos.

Como se verá en la parte de experimentación, para el conjunto global de consultas propuesto para la tarea de recuperación de imágenes de Wikipedia durante las ediciones 2010 y 2011 del *ImageCLEF*, la estrategia LSMF propuesta obtiene los mejores resultados (según MAP) de entre todos los grupos participantes, únicamente casi igualados por una aproximación similar propuesta paralelamente en la edición de 2011 (Csurka, Clinchant and Popescu, 2011).

En resumen, la técnica de fusión multimedia propuesta (LSMF) consigue explotar la complementariedad y aprovecharse de la colaboración entre la información textual y visual presente en las imágenes, siguiendo una estrategia asimétrica de recuperación basada en la mayor carga semántica de la modalidad textual, que permite mejorar los resultados monomodales tanto de TBIR como de CBIR, ayudando a resolver el problema de la brecha semántica (*semantic gap*) tan presente en la tarea de recuperación multimedia de imágenes.

5.5 Contribuciones colaterales a LSMF

Durante el desarrollo de este trabajo, y como parte de la experimentación dedicada al análisis y evaluación de la técnica LSMF, se han analizado otros aspectos relacionados que han desembocado en un número de contribuciones adicionales, como se verá en el Capítulo 7. Cabe destacar el enriquecimiento textual de las anotaciones asociadas a las imágenes haciendo uso de la Wikipedia como recurso externo, el cual mejora significativamente la calidad de la recuperación basada en texto, como se muestra en el apartado 7.1.2.

Se analiza también (apartados 7.2.2 y 7.3.6) el rendimiento de CBIR cuando se lleva a cabo en base a grupos de descriptores visuales independientes, en comparación con cuando se hace con todos ellos de manera conjunta, tanto de la recuperación visual monomodal, como dentro del esquema de fusión multimedia LSMF.

Otro aspecto analizado es el de la influencia de la normalización de las listas de resultados previa a la fusión multimedia. En el apartado 7.3.5 se llevan a cabo varios experimentos con dos técnicas de normalización de *scores* y se compara su comportamiento con el caso de utilizar los valores originales obtenidos desde cada subsistema monomodal.

El rendimiento del prefiltro textual y de la estrategia completa de LSMF es también analizado en función de la dificultad y la carga visual (*visuality*) de las consultas multimodales, en el apartado 7.4. Esta experimentación se realiza en base a la clasificación de las consultas según las características mencionadas proporcionada por los organizadores de la tarea de recuperación de imágenes de *ImageCLEF*.

Dentro del marco del proyecto Buscamedia, se confirma la mejora de los resultados de recuperación al aplicarse técnicas de fusión multimedia, Cuando, además de la información textual asociada a los objetos multimedia de la colección, se utiliza también información procedente de otras modalidades como audio, imagen o vídeo. Esta información multimodal se obtiene a partir del tratamiento de los diferentes objetos multimedia utilizando técnicas visuales como reconocimiento de objetos en imágenes, subtítulos en vídeos, o transcripciones del audio.

Capítulo 6 Entorno desarrollado para la Recuperación Multimedia de Imágenes

En primer lugar se describe globalmente la arquitectura diseñada e implementada, mostrando los cuatro componentes principales del entorno, que serán analizados detalladamente en los siguientes apartados:

- Recuperación de imágenes basada en texto (TBIR)
- Recuperación de imágenes basada en contenido (CBIR)
- Prefiltro semántico textual
- Módulo de fusión

La mayor parte de las funcionalidades de este sistema de recuperación multimedia (exceptuando la recuperación de tipo visual o CBIR) están incluidas dentro de la herramienta IDRA (Granados Muñoz, García Serrano and Goñi Menoyo, 2009), cuya descripción detallada se encuentra en el Anexo. La segunda versión de esta herramienta es descargable desde SourceForge (http://sourceforge.net/projects/idraproject/).

6.1 Arquitectura

La visión general del entorno puede verse en la Figura 6.1, donde se incluyen los componentes principales del mismo, así como las relaciones entre ellos.

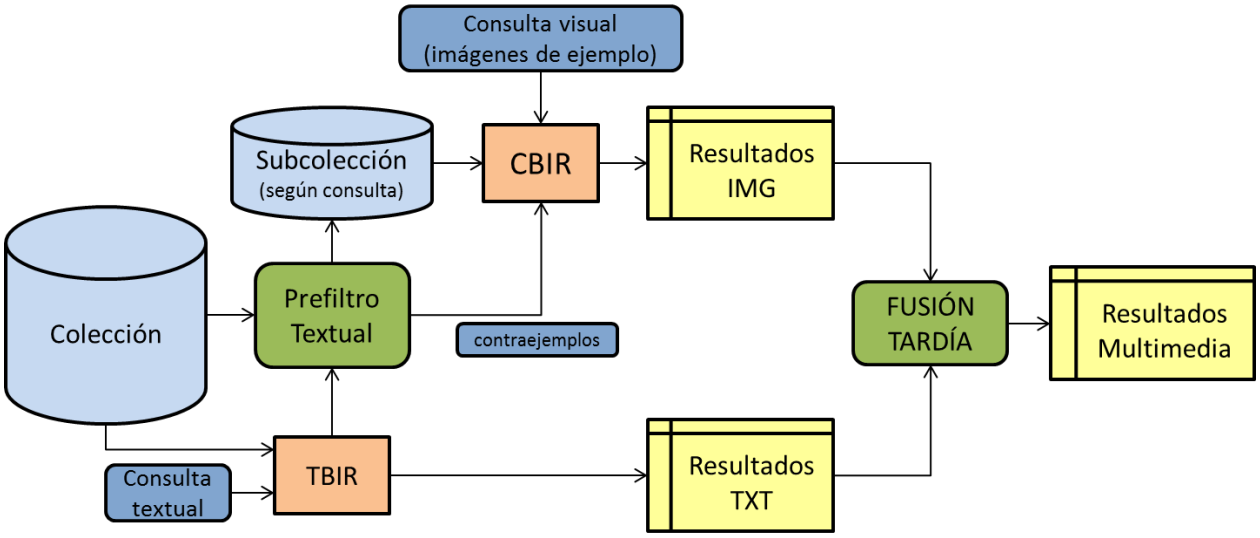

Figura 6.1 Entorno para la Recuperación Multimedia de Imágenes (con LSMF)

El funcionamiento global consiste en los siguientes pasos:

1) tanto el sistema de recuperación de imágenes basado en texto (TBIR) como el basado en el contenido de las imágenes (CBIR) genera su propia lista de resultados a partir de la consulta multimedia y de la información disponible en la colección y en la subcolección.

2) El prefiltro textual selecciona de la colección original un conjunto de imágenes que sirven de contraejemplos para el sistema CBIR.

3) los resultados generados por los sistemas TBIR y CBIR son combinados mediante el módulo de fusión, que generará la lista de imágenes del resultado multimedia.

La herramienta también permite el proceso sin prefiltro textual y, por lo tanto, sin ejemplos negativos de imágenes (contraejemplos) en la parte visual de la consulta multimedia.

6.2 Recuperación de imágenes basada en texto (TBIR)

En la

Figura **6.2** se puede observar la organización de los principales componentes que toman parte en el proceso de recuperación textual, así como el flujo de datos seguido.

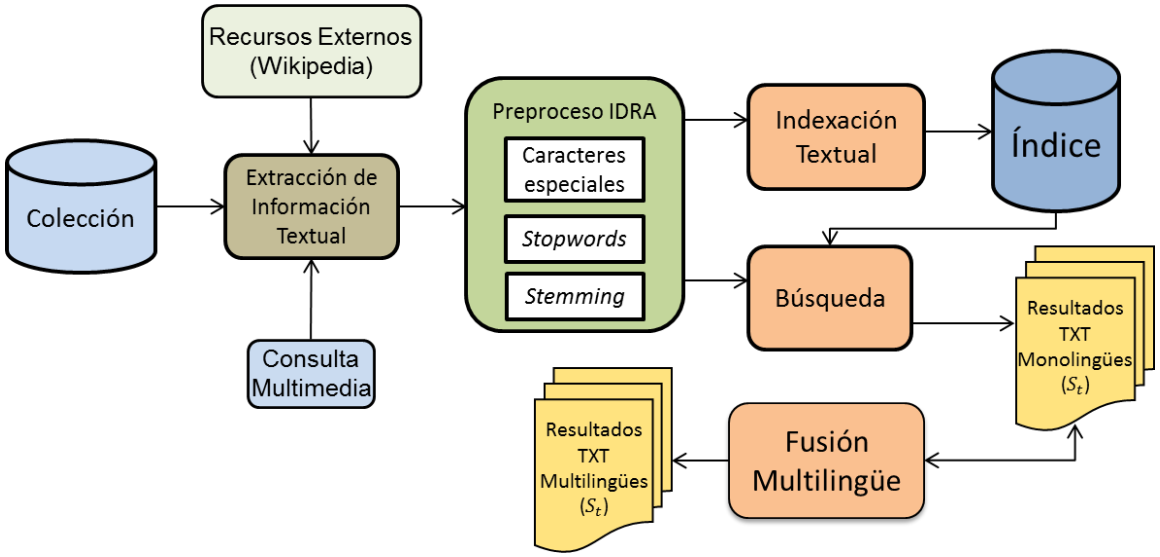

Figura 6.2 Subsistema de Recuperación Textual (TBIR)

El proceso comienza con la extracción de la información textual asociada a las imágenes de la colección (o, en su caso, a la parte textual de las consultas multimedia de evaluación). También pueden utilizarse recursos externos, como por ejemplo Wikipedia, para enriquecer las anotaciones textuales ya disponibles. La información textual extraída es a continuación preprocesada por la herramienta IDRA (tratamiento de caracteres especiales, eliminación de palabras vacías, *stemming*, etc.) para seguidamente ser indexada mediante el módulo de Indexación Textual. Con esto se habrá construido el índice sobre el que se realizarán las consultas del usuario (Búsqueda) también preprocesadas de la misma manera. El proceso de búsqueda y recuperación dará lugar a la lista de resultados textual, que contendrá las decisiones tomadas por el sistema TBIR (conjunto de imágenes recuperadas en base a las características textuales, junto con su valor de relevancia o score textual St), las cuales serán enviadas al módulo de fusión que será el encargado de combinarlas con otras aproximaciones (textuales para experimentos multilingües, o visuales para experimentos multimedia).

- **Extracción de Información Textual**

Este componente se encarga de extraer el texto relevante de los metadatos proporcionados por la colección. En función de la colección con la que se esté experimentando, esta extracción llevará asociada una configuración en particular. Normalmente los metadatos textuales son proporcionados en ficheros XML, con la información textual organizada en los distintos campos (pueden verse los detalles de cada colección en concreto en el Capítulo 4, apartado

4.2). La fase de extracción consistirá en procesar esos ficheros XML para obtener la información textual considerada relevante en cada caso, organizada por idiomas, sin errores de codificación, y lista para su preprocesamiento textual.

Esta misma fase de extracción será aplicada sobre los ficheros con las consultas e, igualmente, dependerá de las particularidades de cada colección y del formato de las consultas.

El esquema de la arquitectura implementada contempla la posibilidad de hacer uso de algún recurso o herramienta externa (Recursos Externos) para enriquecer de algún modo la información textual asociada a las imágenes. Por ejemplo, en el caso concreto de trabajar con la colección de imágenes de Wikipedia, se utilizará la información textual correspondiente a los artículos en los que aparecen las imágenes de la colección para enriquecer las anotaciones disponibles.

- **Preprocesamiento textual con IDRA**

Una vez extraída la información textual asociada a cada elemento de la colección multimedia, esta será preprocesada lingüísticamente como paso previo a su indexación. Las posibles técnicas de preprocesamiento a aplicar, descritas en la sección 2.4.2, son las siguientes:

- Análisis lingüístico y eliminación de caracteres especiales. La información textual disponible es separada en palabras, que serán las candidatas para convertirse en términos de indexación. Además, son eliminados determinados caracteres como tildes, mayúsculas, números, diéresis, etc., dependiendo de las necesidades de la colección.

- Eliminación de *stopwords*. Las palabras consideradas semánticamente vacías serán eliminadas del texto. Al configurar cada experimento, se deben especificar los ficheros contenedores de las listas de *stopwords* correspondientes para cada idioma con el que se trabaje.

- *Stemming*. Las palabras seleccionadas como términos de indexación serán reducidas a su raíz o *stem*, haciendo uso del algoritmo de *stemming Snowball*[26].

[26] http://snowball.tartarus.org

Cada idioma con el que se trabaje en cada experimento concreto utilizará su propia versión del algoritmo.

- **Indexación**

Una vez que se ha extraído la información textual relevante para cada imagen de la colección, y que esta ha sido preprocesada lingüísticamente, se continua con la fase de indexación, lo que dará como resultado la construcción del índice textual. Dentro de la arquitectura propuesta, esta tarea se puede llevar a cabo utilizando dos motores de indexación diferentes (según el experimento). Una opción es utilizar la funcionalidad de indexación de la herramienta IDRA, de implementación propia y fácil de configurar para la ejecución de pruebas concretas. La otra alternativa es utilizar Lucene, proyecto que desarrolla software de búsqueda en código libre, y que es ampliamente utilizado por la comunidad investigadora debido a su buen rendimiento y escalabilidad.

Lucene también permite configurar detalladamente los tipos de preprocesamiento lingüístico aplicados al texto antes de ser indexado pero, como para los experimentos propuestos en esta tesis dicho preprocesamiento es siempre llevado a cabo por la herramienta IDRA, no será necesario utilizar los incluidos en Lucene. El proceso de indexación seguirá, en ambos casos, el modelo del espacio vectorial, detalladamente definido en el apartado 2.1.1.2, con los aspectos comentados relacionados con la normalización.

- **Búsqueda / Recuperación textual**

En esta fase las consultas textuales preprocesadas son lanzadas contra el índice construido. El motor de búsqueda, que puede ser el incluido en la herramienta IDRA o el proporcionado con Lucene, se encargará de generar la lista de imágenes resultado que se obtendrá en función de la información textual. Esta lista estará ordenada según el valor de relevancia o de similitud textual o *score* (S_t) de cada imagen de la colección con respecto a la consulta.

Dentro del modelo del espacio vectorial seguido, la función de similitud utilizada para medir la relevancia de una consulta con respecto a una imagen será la función coseno, que calcula la distancia entre el vector representante de la consulta y el de la anotación textual asociada a cada una de las imágenes de la colección.

En el caso de utilizar la herramienta Lucene la utilización de la medida basada en el coseno no es del todo estricta (McCandless, Hatcher and Gospodnetić, 2010). Por motivos de usabilidad Lucene no calcula exactamente el coseno tal y como se supone en el modelo del espacio vectorial, en el que se normaliza el producto escalar de los vectores ponderados dividiendo entre la norma euclídea de dichos vectores, con lo que se normaliza entre 0 y 1, tal y como se define en el apartado 2.1.1.2. Por el contrario, Lucene utiliza una normalización diferente para la longitud del vector correspondiente al documento, la cual lo convierte en un vector igual o mayor que el vector unidad. En lugar de calcular la norma euclídea para cada documento, lo que supondría un coste computacional muy elevado, Lucene utiliza la longitud del documento para normalizar. Esta longitud es calculada cuando se añade el documento correspondiente al índice, en correspondencia con el número de *tokens* en el mismo. Además, para un eficiente cálculo del valor de relevancia o *score*, la norma euclídea de la consulta se calcula cuando empieza la búsqueda, ya que es independiente del documento para el que se calcula la relevancia. La normalización del vector correspondiente a la consulta proporciona comparabilidad, hasta cierto punto, entre dos o más consultas.

La salida de la fase de búsqueda o recuperación de imágenes será una lista con los resultados obtenidos. La lista incluirá, para cada imagen recuperada, el identificador de la misma, el valor de relevancia o *score* textual con respecto a la consulta introducida, y el orden o *ranking* en la lista de resultados generada.

- **Fusión**

El módulo de fusión multimedia propiamente dicho se describe detalladamente en el apartado 6.4 como un módulo independiente, pero se menciona brevemente aquí la parte de ese subsistema que es utilizada de manera interna por el subsistema textual (sin interacción con los resultados del subsistema visual).

El motivo es que algunos de los experimentos multilingües en los que se trabaja de manera independiente con anotaciones textuales provenientes de diferentes idiomas necesitan diferentes índices. Por lo tanto, la búsqueda en cada uno de estos experimentos monolingües generará una lista de imágenes resultado para cada uno de los idiomas. Para combinar estas listas de resultados monolingües se sigue una estrategia basada en una aproximación de fusión tardía (a nivel de decisiones) con el objetivo de construir una única lista de resultados

(multilingüe). El algoritmo concreto de fusión utilizado (*MaxMerge*) se describe detalladamente en el apartado correspondiente al módulo de fusión.

6.3 Recuperación de imágenes basada en contenido (CBIR)

Este subsistema es el encargado de recuperar el conjunto de imágenes relevantes desde el punto de vista de los descriptores o características visuales extraídas de las imágenes de la colección y de las proporcionadas como parte de la consulta multimedia como ejemplos visuales. La lista de resultados devuelta estará ordenada según la relevancia o similitud visual (S_v) de cada imagen de la colección con la consulta.

El proceso CBIR estará compuesto por dos fases principales: 1) la extracción de características visuales, y 2) el cálculo de la similitud en base a dichas características.

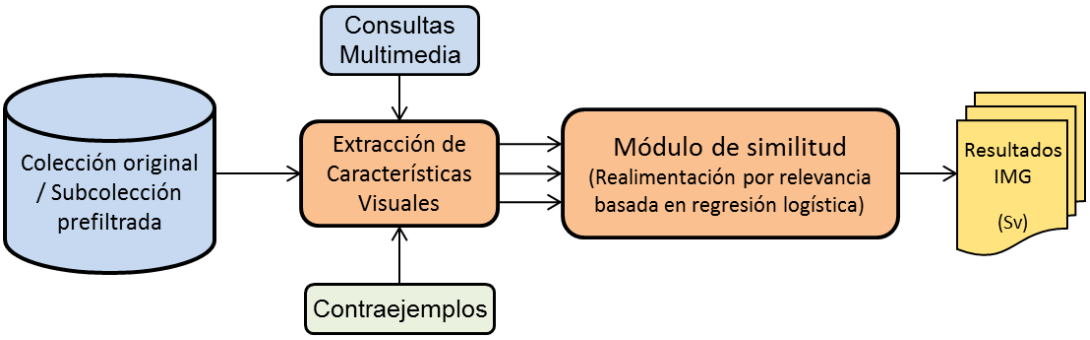

Figura 6.3 Subsistema de Recuperación Visual (TBIR)

La información de entrada al sistema de recuperación visual CBIR será la colección de evaluación y la parte visual de las consultas multimedia, esto es, las imágenes. En el caso del enfoque de fusión multimedia semántica tardía (LSMF) propuesto en esta tesis, el sistema visual trabajará sobre la versión reducida (subcolección) de la colección original gracias a la aplicación del prefiltro textual, y se dispondrá de imágenes negativas proporcionadas por el prefiltro, que servirán de contraejemplos para el algoritmo de recuperación visual.

Se describen a continuación los componentes principales de este subsistema:

- **Extracción de características**

Este componente analiza tanto las imágenes de la colección como las proporcionadas como ejemplo en la parte visual de las consultas multimedia (también de las imágenes

contraejemplo en caso de disponer de ellas), extrayendo de ellas las características visuales o descriptores de bajo nivel con los que serán representadas.

La experimentación llevada a cabo en relación a la recuperación visual (descrita con detalle en el Capítulo 7, apartado 7.2), se realiza haciendo uso de los descriptores visuales CEDD (*Color and Edge Directivity Descriptor*). Se trata de un conjunto de descriptores proporcionados por la organización de la tarea de recuperación de imágenes de Wikipedia del *ImageCLEF*. Estos descriptores, que incluyen más de una característica en un histograma compacto, pertenecen a la familia de los CCD (*Compact Composite Descriptors*, Descriptores Compactos Compuestos). Se componen de 6 zonas de textura, divididas cada una en 24 subregiones que describen cada una un color. La información sobre el color en los descriptores CEDD proviene de 2 sistemas difusos o borrosos que asocian los colores de la imagen con una paleta personalizada de 24 colores. Para extraer información sobre la textura, CEDD utiliza una versión borrosa de los 5 filtros digitales propuestos por EHD MPEG-7. El histograma se normaliza dentro del intervalo [0,1] y para la representación binaria en una cuantificación de 3 bits. La característica más importante de los descriptores CEDD es el haber conseguido muy buenos resultados con varias bases de datos de imágenes (*WANG's, MPEG-7 CCD, UCID43, img(Rummager) and Nister database*) (Chatzichristofis et al., 2010).

La tarea de extracción de características es un proceso costoso, en el que hay que calcular cada una de ellas para todas las imágenes de la colección. El tiempo de extracción para una sola imagen se ha calculado y supone aproximadamente 0,69 segundos. Por lo tanto, la extracción de las características visuales de una colección como la proporcionada en *ImageCLEF*, utilizada en la fase de experimentación de esta tesis (237.434 imágenes), llevará aproximadamente dos días de computación.

- **Módulo de similitud**

El módulo de cálculo de similitud visual recibe como entrada las características visuales de bajo nivel (color, forma, textura, etc.) tanto de las imágenes de la colección como de las imágenes proporcionadas como ejemplo dentro de la parte visual de las consultas multimedia. La salida será asociar a cada imagen de la colección de entrada un valor de semejanza o similitud visual (S_v) con la consulta (esto es, la imagen o grupo de imágenes dadas como

ejemplo). El módulo de similitud también puede tener en cuenta ejemplos de imágenes negativas (contraejemplo) como entrada a su algoritmo de cálculo de similitud.

En el conjunto de experimentos visuales desarrollados en el apartado 7.2.1, se trabaja con varios algoritmos de recuperación visual: automático, de expansión de la consulta, y de realimentación por relevancia. Estos algoritmos son brevemente descritos en dicho apartado. A continuación se define con detalle el algoritmo utilizado en la configuración final del sistema CBIR, el algoritmo de realimentación por relevancia basado en regresión logística (León et al., 2007), implementado por el grupo de investigación de la UV con el que se colabora para la parte visual de los experimentos multimedia.

El algoritmo calcula la probabilidad o *score* (S_v) de que una imagen pertenezca al conjunto de imágenes buscadas, y modela la función *logit* ($logit(p) = \log(p) - \log(1-p)$) de esta probabilidad como la salida de un modelo lineal generalizado cuyas entradas son las características visuales de bajo nivel de la imagen. El algoritmo necesita ejemplos y contraejemplos visuales (imágenes positivas y negativas). Las imágenes positivas serán las proporcionadas como ejemplo en cada consulta o *topic* multimedia de evaluación. Como las colecciones de evaluación de *ImageCLEF* que se utilizan en los distintos experimentos no proporcionan ejemplos negativos de imágenes para las consultas multimedia que proponen, estas deben obtenerse de alguna manera para el correcto funcionamiento del algoritmo. Dentro de la propuesta LSMF, la estrategia seguida para seleccionar los M ejemplos negativos para el algoritmo de regresión consiste en seleccionar J imágenes al azar de entre las que no superen el prefiltro semántico textual, y ordenarlas en base a la distancia euclídea. Las M últimas imágenes de esta lista serán las elegidas como imágenes negativas.

Una vez que el sistema CBIR dispone tanto de las imágenes de ejemplo (positivas) como de las de contraejemplo (negativas), se ejecuta el algoritmo de realimentación por relevancia (*relevance feedback*) basado en regresión logística:

- Se considera que la variable aleatoria Y representa la evaluación por parte del usuario ($Y = 1$ significa que la imagen es relevante, $Y = 0$ no relevante).

- Cada imagen de la colección está descrita con sus descriptores de bajo nivel tras la fase de extracción de características, de tal manera que la imagen j-ésima tiene asociado su vector de características k-dimensional correspondiente x_j.

- Los datos serán los pares (x_j, y_j), con $j = 1 \dots n$, donde *n* es el número total de imágenes, x_j es el vector de características, e y_j la evaluación de la imagen $(1 = positivo, 0 = negativo)$.

- El vector de características x es conocido para todas las imágenes y se trata entonces de predecir el valor asociado Y. Para ello se utiliza una regresión logística para $P(Y = 1|x)$, esto es, la probabilidad de que $Y = 1$ (imagen evaluada positivamente) dado que el vector de características x está relacionado con una combinación lineal del vector de características por medio de la función $logit$.

- Para una variable Y de respuesta binaria y p variables explicativas $x_1, \dots, x_p$, el modelo para $\pi(x) = P(Y = 1|x)$ en los valores de $x = (x_1, \dots, x_p)$ de predictores, es $logit[\pi(x)] = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p$, donde $logit[\pi(x)] = \ln\left(\pi(x)/(1 - \pi(x))\right)$. Los parámetros del modelo se obtienen mediante la maximización de la función de probabilidad dada por:

$$l(\beta) = \prod^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}$$

- El estimador de máxima verosimilitud (MLE, *maximum likelihood estimator*) del vector de parámetros β se calcula utilizando un método iterativo.

La dificultad es tener que ajustar un modelo de regresión general, en el que se toman en cuenta el conjunto completo de variables, ya que el número de imágenes seleccionadas *k* (ejemplos positivos y negativos) normalmente es menor que el número de características *p* ($k < p$). Por eso el modelo de regresión ajustado tendrá tantos parámetros (β) como la cantidad de datos, y muchas variables relevantes podrían no ser consideradas. Con el fin de resolver este problema, la propuesta (Granados et al., 2011) consiste en ajustar diferentes modelos de regresión más pequeños: cada modelo considerará sólo un subconjunto de

variables de características de la imagen relacionadas semánticamente. En consecuencia, cada sub-modelo asociará una probabilidad de relevancia diferente a una determinada imagen x, y habrá que combinarlas con el fin de clasificar las imágenes de acuerdo a las preferencias del usuario. Este problema ha sido resuelto por medio de un operador OWA (*Ordered Weighted Averaged* (Yager, 1988)), que combinará las decisiones tomadas por cada sub-modelo de regresión generando un valor de similitud o *score* final (S_v) para la imagen x.

Se define a continuación el procedimiento general descrito en forma de algoritmo:

ENTRADAS

- Sea I_i el conjunto de imágenes en la colección, $i \in \{1, \dots, N\}$.
- Sea $ZP^p = (x_1^p, x_2^p, \dots, x_k^p, y_n)$ el conjunto de vectores de características en R^kde las imágenes positivas seleccionadas (de tamaño L) más su evaluación ($y_p \leftarrow 1$).
- Sea $ZN^p = (x_1^n, x_2^n, \dots, x_k^n, y_n)$ el conjunto de vectores de características en R^kde las imágenes negativas seleccionadas (de tamaño K) más su evaluación ($y_n \leftarrow 1$).
- Sea $g_o$el número de grupos de características, $g_o \leq k$.

Estimación de parámetros por MLE (estimador de máxima verosimilitud)

- Inicialización:

$$ini \leftarrow 1 \qquad f_i \leftarrow [k/g_o]$$

- **FOR** $r = 1, \dots, g_o$**DO**
  - construir la matriz de datos $W_{Mx[\frac{k}{g_o}]} = (WP; WN)$ con ejemplos positivos y negativos ($M = L + K$) para cada grupo de características:

$$WP = (x_{ini}^p, \dots, x_{f_i}^p, y_p) \qquad WN = (x_{ini}^n, \dots, x_{f_i}^n, y_n)$$

  - estimar los parámetros $\overline{\mu} = (\alpha, \beta_1, \dots, \beta_{(\frac{k}{g_o})})$ para los datos W mediante el estimador MLE, lo que implica calcular el máximo de:

$$l(\mu) = \prod_1^M \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}$$

donde $ln \frac{\pi(x)}{1-\pi(x)} = 1 + \alpha + \beta_1 + \cdots + \beta_{(\frac{g_o}{k})}$

  - predecir la probabilidad $\pi_r(I_i)$ para cada imagen de la colección los parámetros de estimación $\overline{\mu}$.
  - actualizar:

$$ini \leftarrow f_i + 1 \qquad f_i \leftarrow r\ x\ [k/g_o]$$

- **END**

SALIDA

- Se utiliza un operador OWA para fusionar el vector de probabilidades $\pi(I_i) = (\pi_1(I_i), \dots, \pi_{g_o}(I_i))$ en un único score

$$\pi_{f_i}(I_i) = OWA(\pi(I_i))$$

Figura 6.4. Algoritmo de realimentación por relevancia basado en regresión logística

La salida final del sistema de recuperación visual será la lista de imágenes de la colección ordenadas según el valor o *score* de semejanza de cada una de ella (S_v entre 0 y 1) con respecto a las imágenes de ejemplo proporcionadas como parte de la consulta multimedia, y teniendo en cuenta los contraejemplos (imágenes negativas) facilitados por el prefiltro textual.

En cuanto al tiempo estimado para calcular la similitud visual, trabajando con la colección de imágenes de Wikipedia de *ImageCLEF*, entre una consulta y cada una de las imágenes de la colección (un total de 237.434) es de aproximadamente 10 horas en los experimentos realizados.

6.4 Módulo de Fusión

Este módulo incorpora varios algoritmos de fusión a nivel de decisiones, implementados para llevar a cabo la combinación de diferentes listas de resultados. Estas listas pueden ser las obtenidas desde los sistemas de recuperación monomodales (para generar un resultado multimedia), o también las generadas por experimentos textuales monolingües (cuya combinación resultará en una lista de resultados multilingüe).

Para el caso concreto de la tarea de recuperación multimedia de imágenes, se combinarán las listas generadas por los sistemas de recuperación textual y visual, las cuales incluirán las imágenes recuperadas para cada una de las consultas, junto con el valor de relevancia obtenido en cada caso (S_t desde el subsistema textual, y S_v desde el visual). Estas listas seguirán el formato TREC, descrito en el apartado 4.3.3. La salida proporcionada por el módulo de fusión será una única lista de imágenes resultado junto con su valor de relevancia o *score* fusionado.

Como se ha visto en el estado del arte (apartado 3.4), los métodos de fusión multimedia tardía (*late fusion*) pueden estar basados en los *scores* monomodales, en las posiciones obtenidas por los elementos multimedia en las listas de resultados (*ranking*), o en ambos tipos de información. Los algoritmos disponibles en este módulo están implementados cada uno de ellos con objetivos concretos, y pueden estar basados en el *score*, en el *ranking*, o en ambos.

Como paso previo a esta fusión tardía puede llevarse a cabo una fase de normalización de los valores de relevancia (*scores*) obtenidos en cada una de las listas de resultados. En el apartado 7.3.4 se analizan un conjunto de experimentos en relación a esta fase de normalización.

Los algoritmos de fusión tardía, que se describen a continuación, suponen el escenario de la recuperación multimedia de imágenes, en el que se dispone de las listas de resultados textual y visual:

MaxMerge. Este algoritmo construye la lista de resultados fusionada seleccionando aquellas imágenes que tengan un valor de similitud o *score* mayor, independientemente de si este ha sido obtenido desde el subsistema textual (S_t) o desde el visual (S_v). Por lo tanto, el *score* en la lista final fusionada (S_f) de una imagen i en relación a una determinada consulta Q será:

$$S_f(i, Q) = max(S_t, S_v)$$

El funcionamiento de este algoritmo es equivalente al de *combMAX*, visto en la revisión del estado del arte sobre técnicas de fusión, concretamente dentro de las correspondientes a las funciones de agregación de *scores* (apartado 3.4.1).

FilterN. El objetivo de este algoritmo de fusión es eliminar de la lista de resultados textuales aquellas imágenes que no aparezcan entre las N primeras posiciones de la lista visual. La idea es confiar inicialmente en las decisiones tomadas en base a las anotaciones textuales (S_t), pero eliminando aquellas imágenes que según el subsistema visual no tengan una semejanza suficiente (S_v) y no se encuentren entre las N primeras del ranking visual. La lista de resultados fusionada se construirá en función del nuevo *score* (S_f) calculado tras la combinación multimedia:

$$S_f(i, Q) = I_{\{i \in topN_t(Q)\}} \cdot S_t(i, Q)$$

donde:

$topN_t(Q)$: conjunto de las N imágenes resultado más similares para la consulta Q, en base a la recuperación textual

$$I_{\{A\}} = \begin{cases} 1, & preprosición\ A\ es\ verdadera \\ 0, & e.o.c. \end{cases}$$

El valor de N que marcará la posición en la lista visual a partir de la cual se eliminan las imágenes de la lista textual es parametrizable. En el conjunto de experimentos del Capítulo 7

que involucran a este algoritmo de fusión, se analiza el comportamiento del mismo para diferentes valores de este parámetro.

Las imágenes con un valor de relevancia o *score* tras la fusión (S_f) igual a 0, no serán incluidas en la lista de resultados definitiva tras el proceso de fusión multimedia.

Con este algoritmo se trata de limpiar la lista de resultados textuales, en base en las posiciones ocupadas por las imágenes en la lista de resultados visual, para generar una lista multimodal de resultados más precisa y relevante para las consultas lanzadas.

Product. La lista multimedia final será construida en base al producto de los *scores* procedentes de los subsistemas monomodales (S_t y S_v) para cada imagen.

$$S_f(i,Q) = S_t(i,Q) \cdot S_v(i,Q)$$

Este algoritmo de fusión multimedia tardía es equivalente a la técnica *combPROD* vista en la revisión del estado del arte sobre estrategias de fusión basadas en agregación de *scores* (apartado 3.4.1).

OWA. Tomando como entradas los valores de similitud obtenidos por los subsistemas monomodales (S_t y S_v) para cada imagen, este algoritmo hace uso de distintas configuraciones del operador de media ponderada ordenada OWA (*Ordered Weighted Averaging* (Yager, 1988)) con el fin de combinar los resultados de ambos subsistemas. Este operador transforma un número finito de entradas (para el caso de la recuperación multimedia de imágenes serán dos entradas) en una única salida. Con el operador OWA no se asocia ningún peso a una ninguna entrada en particular, sino que la magnitud relativa de la entrada (mayor o menor *score*) decide qué peso le corresponde a cada una de ellas. Esto introduce la no linealidad en el proceso de agregación (Carlsson and Fuller, 2002), lo que puede resultar una ventaja cuando no se conoce a priori qué subsistema (textual o visual) proporcionará la mejor información.

Formalmente, un operador OWA de dimensión n es una aplicación $F: R_n \rightarrow R_n$ que tiene un vector de ponderaciones asociados $W = [w_1, w_2, \dots, w_n]$ tal que:

$$w_i \in [0,1], 1 \leq i \leq n$$

$$\sum_{i=1}^{n} w_i = 1$$

$$F(x_1, x_2, \dots, x_n) = \sum_{k=1}^{n} w_k \, x_{jk}$$

Con los ejemplos que se muestran a continuación puede apreciarse la generalidad de esta técnica de agregación de scores al poder obtenerse un gran número de operadores según la elección de los pesos. Se muestran los casos que simularían el máximo, el mínimo, y la media aritmética:

- $W = [1,0,\dots,0]^T \rightarrow F(x_1, \dots, x_n) = \min(x_1, \dots, x_n)$
- $W = [0,0,\dots,1]^T \rightarrow F(x_1, \dots, x_n) = \max(x_1, \dots, x_n)$
- $W = \left[{}^1/_N, {}^1/_N, \dots, {}^1/_N\right]^T \rightarrow F(x_1, \dots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$

Los operadores OWA están delimitados por el operador *max* (OR) y por el *min* (AND). Por esto razón, en (Filev and Yager, 1997) se introduce la medida *orness*, que permite definir el grado de optimismo (o esperanza positiva) de la agregación basada en el operador OWA:

$$orness(W) = \frac{1}{n-1}\sum_{i=1}^{n}((n-i)w_i)$$

Esta medida *orness*, que puede tomar valores entre 0 y 1, indica el grado en que la agregación se asemeja a una operación OR (*max*) o, como se ha comentado antes, su grado de optimismo:

- $orness(W = [1, 0, \dots, 0]^T) = 1$ (el score de salida tras la fusión será el mayor score mono-modal de entrada) → grado de optimismo máximo
- $orness(W = [0, 0, \dots, 1]^T) = 0$ (el score de salida tras la fusión será el menor score mono-modal de entrada) → grado de optimismo mínimo
- $orness\left(W = \left[{}^1/_n, {}^1/_n, \dots, {}^1/_n\right]^T\right) = 0.5$ (el score de salida tras la fusión será la media aritmética de los scores mono-modales de entrada) → grado de optimismo neutro

Aunque en este trabajo se definirán los operadores OWA utilizados en los diferentes experimentos en base a la medida *orness*, cabe mencionar que también existe la medida $andness$ que define la esperanza negativa (complementaria al $orness$) de un operador OWA:

$$andness(W) = 1 - orness(W)$$

En resumen, los operadores con muchos pesos cerca del más alto serán similares a un operador tipo OR ($orness(W) \leq 0.5$), mientras que los operadores con la mayoría de pesos inferiores serán similares a uno de tipo AND ($orness(W) \geq 0.5$) (Fernández Salido and Murakami, 2003).

Como podrá verse en la sección correspondiente a la experimentación (Capítulo 7), los pesos de agregación utilizados cubren todo el rango de valores que van desde el mínimo hasta el máximo con los diferentes pesos utilizados de $ornesss$. Por ejemplo, este trabajo contempla dos *scores* de entrada (uno calculado por el subsistema textual, S_t, y otro por el visual, S_v), un $orness = 0.3$ significará que se le da un peso de 0.3 al valor de probabilidad o *score* más alto, y un peso de 0.7 a la probabilidad o *score* más bajo de las entradas.

Enrich. Trabaja a partir de las listas de resultados generadas desde los subsistemas textual y visual, que contienen las imágenes recuperadas ordenadas en función del valor de relevancia o *score* obtenido en cada caso (S_t y S_v). Este algoritmo toma la lista de imágenes resultado del subsistema textual como principal, y la del visual como secundaria o de soporte. El objetivo es enriquecer la lista de decisiones textuales con las visuales, en función de la posición y del *score* de las imágenes recuperadas por el subsistema CBIR. El valor de relevancia o *score* de las imágenes tras la fusión (S_f) se calcula de la siguiente manera:

$$S_f(i,Q) = S_t(i,Q) + (\frac{S_v(i,Q)}{R_v(i,Q)+1})$$

donde:

$S_f(i,Q)$: *score* o valor de relevancia tras la fusión multimedia de la imagen i con respecto a la consulta Q.

$S_t(i,Q)$: *score* o valor de relevancia decidido por el subsistema textual de la imagen i con respecto a la consulta Q.

$S_v(i,Q)$: *score* o valor de relevancia decidido por el subsistema visual de la imagen i con respecto a la consulta Q.

$R_v(i, Q)$: posición o *ranking* de la imagen i en la lista de resultados visuales para la consulta Q.

Las imágenes que aparecen en la lista de soporte, pero no en la principal, se añaden al final de la lista fusionada. En estos casos, los valores de relevancia se normalizan de acuerdo al valor más bajo en la lista resultante de la fusión multimedia.

6.5 Comentarios finales

Los sistemas de recuperación de imágenes aquí descritos (tanto monomodales como multimedia) serán los evaluados en los distintos conjuntos de experimentos planteados en el Capítulo 7. Respecto a la fusión de resultados tardía, se han planteado e implementado cinco algoritmos diferentes, cuyo comportamiento será analizado y comparado en el mismo capítulo.

La fusión de resultados monolingües se realiza también con uno de estos algoritmos, dependiendo de las pruebas y la colección, justificando la selección por los resultados del conjunto de "entrenamiento" siempre que esté disponible.

Capítulo 7 Experimentación

Se incluyen en primer lugar los experimentos llevados a cabo en relación a la recuperación de imágenes basada en texto (TBIR), y a continuación los correspondientes a la recuperación basada en el contenido visual de las imágenes (CBIR). El tercer apartado está dedicado a la experimentación referente a la fusión multimedia entre los resultados textuales y visuales, dentro del cual se evalúa la propuesta principal de este trabajo (LSMF). Después se incluye un análisis de la influencia en los resultados del tipo de consultas tratadas, en función de su complejidad y su carga visual. Finalmente, se describe la experimentación llevada a cabo dentro del marco del proyecto de investigación Buscamedia.

7.1 Recuperación de imágenes basada en texto (TBIR)

En este apartado se mostrarán los experimentos realizados dentro del enfoque basado en texto para la recuperación multimedia de imágenes (TBIR). En ellos la única información que se tendrá en cuenta será la de tipo textual, y no se utilizará en ningún caso la información visual.

Los experimentos están orientados a optimizar el procesamiento de la información textual asociada a las imágenes, con el fin de mejorar el rendimiento de la recuperación.

El análisis del rendimiento de las diferentes configuraciones se hará en base a los resultados obtenidos en los diferentes experimentos desarrollados, utilizando la medida de evaluación MAP (*mean average precision*) como indicador principal, ya que es la medida habitualmente utilizada en CLEF. Esta medida es utilizada para tareas de recuperación debido a sus capacidad de evaluar el número de resultados relevantes recuperados (en relación al total de relevantes existentes), teniendo además en cuenta la posición de dichos resultados en el

ranking. En muchos casos se analizarán también los valores de precisión a bajo nivel (*early precision*), esto es, el rendimiento del sistema TBIR según las primeras imágenes recuperadas como respuesta a las búsquedas que, en la mayoría de los casos, son las únicas en las que se interesa un usuario corriente.

Los resultados finales comparados con las aproximaciones seguidas por otros grupos participantes en la tarea de recuperación de imágenes de Wikipedia del *ImageCLEF* (evaluando conjuntamente con las consultas propuestas en las ediciones 2010 y 2011), son mejores. Además, la técnica de recuperación aplicada no utiliza técnicas de expansión de consulta (*query expansion*) ni de realimentación por relevancia (*relevance feddback*), cosa que sí hacen algunas de las soluciones presentadas por otros grupos participantes, con una mayor carga computacional.

La colección de imágenes anotadas utilizada para esta parte de la experimentación es la proporcionada en la tarea de recuperación de imágenes de Wikipedia en la edición del *ImageCLEF* 2010 (utilizada también en la edición 2011). Junto con las imágenes se facilita un conjunto de consultas multimodales o *topics* (70 en 2010 y 50 en 2011), y sus correspondientes juicios de relevancia. Una descripción detallada de la colección puede encontrarse en el apartado 4.2.1.3 de este trabajo.

En algunas de las tablas que muestran los resultados obtenidos por los diferentes experimentos ejecutados, se incluye una columna indicando la "Mejora" de un experimento con respecto a otro. Este valor hace referencia a la mejora relativa, expresada en términos de porcentaje, y es calculada en base a la siguiente función:

$$Mejora(x, y) = \frac{y - x}{x} \cdot 100$$

El porcentaje calculado se refiere a la mejora relativa del segundo valor (y) con respecto al primero (x). Por ejemplo, si un sistema obtiene un $MAP = 0.30$ y otro un $MAP = 0.45$, la mejora relativa del segundo con respecto al primero sería del 50%.

Los experimentos de esta parte textual están organizados de la siguiente manera:

1) se analizan aspectos relacionados con el preprocesamiento (listas de stopwords, uso de *stemming*, identificación de entidades nombradas, herramienta de indexación y

búsqueda, o selección de metadatos), con el objetivo de definir la configuración más apropiada.

2) se explora la conveniencia o no de enriquecer la información textual asociada a las imágenes haciendo uso de Wikipedia como recurso externo.

3) se analiza la mejor manera de combinar la información textual multilingüe, ya que las anotaciones pueden estar disponibles en más de un idioma.

Toda la experimentación se lleva a cabo inicialmente sobre la colección de imágenes de Wikipedia facilitada por los organizadores de la tarea de recuperación del *ImageCLEF* 2010, haciendo uso del conjunto de consultas o *topics* proporcionados en esa misma edición (un total de 70). A continuación se muestra la configuración final del sistema de recuperación textual, definida en base a los resultados y conclusiones obtenidos en los experimentos previos, y se confirmará el buen funcionamiento de las decisiones tomadas evaluando de nuevo la configuración sobre la misma colección pero, en este caso, haciendo uso del conjunto de *topics* propuestos para la edición de 2011 (un total de 50). Finalmente, se realiza una comparación de los resultados con distintas aproximaciones propuestas por otros grupos de investigación participantes en la misma tarea de recuperación (tanto en edición de 2010 como en la de 2011).

7.1.1 Preprocesamiento

Para analizar la influencia de cada uno de los tipos de preprocesamiento se realizan diversos experimentos en los que una misma configuración del subsistema textual es utilizada, variando únicamente aquel componente o técnica en particular cuyo comportamiento se esté evaluando (*stemming*, metadatos, *stopwords*, etc.).

Listas de stopwords

Una buena selección de la lista de palabras vacías (o *stopwords*) puede resultar crucial para el buen funcionamiento de un sistema de recuperación textual, como queda descrito en el apartado 2.1.2. La fase de eliminación de *stopwords* es interesante para el rendimiento de un sistema de recuperación textual (Schauble, 1997) basado en un modelo vectorial. Se evalúa el funcionamiento del subsistema textual propuesto utilizando 2 listas diferentes de *stopwords*:

1) *Stopwords* por defecto de la herramienta IDRA. Se trata de una lista de palabras de parada compuesta por 1.103 términos, incluyendo números y construcciones típicas del idioma (como *ain't*, *didn't* o *doesn't* para el inglés). Esta lista fue construida por un lingüista experto en recuperación de información.

2) *Stopwords* por defecto de Lucene. Lista de palabras de parada incluida como parte del analizador básico de Lucene (*StandardAnalyzer*), dependiente del idioma.

En la siguiente tabla se muestran los resultados obtenidos para los dos experimentos realizados utilizando tanto la lista de *stopwords* de IDRA como la de Lucene (en inglés). La diferencia entre los experimentos 1 y 2 es la aplicación o no de *stemming* a los textos.

Tabla 7-1. Comparación entre listas de *stopwords*

	Lista SW	**MAP**	**Mejora**
Experimento1*	Lucene	0,1752	3,94%
	IDRA	0,1821	
Experimento2*	Lucene	0,2096	5,20%
	IDRA	0,2205	

(*) Mejoras estadísticamente significantes

Analizando lo ocurrido en ambos experimentos se puede observar que la utilización la lista de palabras proporcionada por la herramienta IDRA proporciona resultados ligeramente mejores que cuando se utilizan las *stopwords* por defecto de Lucene.

Stemming

Se realizan tres experimentos para evaluar la conveniencia o no de aplicar técnicas de *stemming* durante la fase de preprocesamiento textual. Los experimentos se diferencian en la lista de *stopwords* utilizada y en la herramienta de indexación/recuperación.

Tabla 7-2. Comparación sobre *stemming*

	Stemming	**MAP**	**Mejora**
Experimento1*	NO	0,1821	21,09%
	SI	0,2205	
Experimento2*	NO	0,1752	19,63%
	SI	0,2096	
Experimento3*	NO	0,1809	15,87%
	SI	0,2096	

(*) Mejoras estadísticamente significantes

En todos los casos analizados los resultados obtenidos con la utilización del *stemming* en la fase de preprocesamiento, tanto de las anotaciones textuales asociadas a las imágenes de la colección como de la parte textual de las consultas, son superiores. En este caso, las mejoras son considerables: alrededor del 20%. Por lo tanto, los experimentos indican que el uso de *stemming* es recomendable cuando se afronta la tarea de recuperación de imágenes en base a sus anotaciones textuales. Los tres experimentos son realizados sobre la parte de la colección disponible en inglés, ya que la colección está mayoritariamente en dicho idioma (140.683 imágenes, casi un 60%).

Herramienta

Lo que aquí se compara es el rendimiento de la recuperación cuando se realiza utilizando la herramienta IDRA y cuando se hace en base a la configuración estándar de Lucene. Se plantean dos experimentos. En el primero (Experimento1), la configuración de los sistemas comparados es idéntica (incluso utilizan la misma lista de *stopwords*, la de IDRA) salvo el parámetro que hace referencia a la herramienta de indexación y búsqueda utilizada. En el segundo (Experimento2), cada sistema comparado utiliza, además de su propia herramienta (IDRA o Lucene), su propia lista de *stopwords*. La siguiente tabla muestra los resultados obtenidos:

Tabla 7-3. Comparación de herramientas

	Herramienta	Stopwords	MAP	Mejora
Experimento1*	IDRA	IDRA	0,1809	0,66%
	Lucene		0,1821	
Experimento2*	Lucene	Lucene	0,1752	3,25%
	IDRA	IDRA	0,1809	

(Los resultados mostrados no superan el test de significancia estadística)

Cuando ambos sistemas de recuperación utilizan su propia herramienta de indexación y búsqueda (incluyendo sus propias listas de *stopwords*), los valores de MAP obtenidos muestran un mejor funcionamiento de la herramienta IDRA, con una mejora relativa en los resultados del 3,25%. En cambio, cuando los parámetros de configuración son exactamente los mismos, incluido el de la lista de *stopwords* empleada, la herramienta Lucene mejora ligeramente a IDRA (0,66%). En cualquier caso, los resultados con ambas herramientas son bastante parecidos y los análisis estadísticos realizados muestran que las diferencias no son significativas.

Metadatos

De entre los campos que contienen información textual dentro de los ficheros de metadatos asociados a cada una de las imágenes de la colección, descritos en el apartado 4.2.1.3, se trata de decidir cuáles de ellos aportan información relevante que sea útil para la tarea de recuperación. Se muestran aquí dos experimentos que tienen el objetivo de aclarar la conveniencia o no de utilizar el texto disponible en el campo <*comment*> general, un campo que se incluye fuera de la información básica identificada por idiomas, y que otros trabajos deciden no utilizar argumentando que la información proporcionada se solapa con la ya dada para cada idioma (Zagoris, Arampatzis and Chatzichristofis, 2010). Como se verá en los experimentos mostrados a continuación, la información sí resulta útil en base a los resultados obtenidos.

Los experimentos 1 y 2 mostrados en la tabla siguiente comparan configuraciones que solo se diferencian en la utilización o no del campo mencionado.

Tabla 7-4. Comparación sobre el uso de distintos metadatos

	<*comment*> general	**MAP**	**Mejora**
Experimento1*	NO	0,1464	19,67%
	SI	0,1752	
Experimento2*	NO	0,1370	32,04%
	SI	0,1809	

(*) Mejoras estadísticamente significantes

En ambas comparaciones (en el Experimento1 se utiliza la herramienta Lucene y en el Experimento2 IDRA) los resultados obtenidos son sensiblemente mejores cuando se incluye la información textual proporcionada por el campo <*comment*> general de los metadatos. El aumento en cuanto a la medida de evaluación MAP supone una mejora relativa de casi el 20% en el primer experimento y de más del 30% en el segundo.

7.1.2 Enriquecimiento textual utilizando Wikipedia

Se plantean en este apartado un conjunto de experimentos dedicados a evaluar la posibilidad de enriquecer las anotaciones de las imágenes de la colección en base a la información textual presente en los artículos de Wikipedia.

Tras el análisis del tipo de información proporcionada por los ficheros que contienen el texto de los artículos de Wikipedia, se decide extraer aquella información textual presente en los campos correspondientes al título y a la categoría del artículo en el que aparece cada imagen (<*title*> y <*categories*>). Este texto será incluido como parte de las anotaciones textuales de la imagen correspondiente para su posterior indexación. El motivo para seleccionar los campos mencionados es que estos dos campos se corresponden con información general precisa sobre el tema del artículo, con el fin de no introducir información ruidosa. Dado que los artículos de Wikipedia pueden ser de una extensión variable, bastante amplia en muchos casos, la información textual en ellos contenida puede tratar diversos aspectos (siempre dentro del tema principal) y, por lo tanto, dicha información no siempre estará directamente relacionada con la imagen a enriquecer.

Se llevan a cabo tres parejas de experimentos, una para cada idioma disponible en la colección (inglés, francés y alemán) en los que se compara el rendimiento del sistema de recuperación de imágenes cuando éste utiliza la información textual extraída de Wikipedia, y

cuando no. Los valores de precisión MAP obtenidos, junto con las mejoras conseguidas para cada idioma, se muestran en la siguiente tabla:

Tabla 7-5. Enriquecimiento textual con Wikipedia

	Idioma	Enriquecimiento Wikipedia	MAP	Mejora
Experimento1*	FR	NO	0,1032	21,32%
		SI	0,1252	
Experimento2*	DE	NO	0,1123	10,69%
		SI	0,1243	
Experimento3*	EN	NO	0,2096	9,88%
		SI	0,2303	
Experimento4*	ALL	NO	0,1654	54,78%
		SI	0,2560	

(*) Mejoras estadísticamente significantes

La mejora alcanzada gracias al enriquecimiento textual propuesto consigue unos valores de precisión (MAP) superiores a todos los obtenidos por los experimentos textuales monolingües de los grupos presentados en la competición del *ImageCLEF* (edición 2010) para el inglés y el alemán (7 y 2 grupos participantes respectivamente), y queda en segunda posición para el francés (4 grupos participantes). Estos permiten mejorar el rendimiento del sistema global de recuperación de imágenes basada en texto, con una influencia directa en los resultados multimedia fusionados finales.

7.1.3 Entidades nombradas

La experimentación relacionada con la reconocimiento de entidades nombradas (NER) se realiza utilizando la colección de imágenes de Wikipedia del *ImageCLEF* 2011 por un lado, y la colección IAPR TC-12 (descritas en el apartado 4.2.1). Se trata de dos grupos de experimentos independientes y distintos, que son descritos a continuación. El objetivo en ambos casos es analizar la conveniencia o no de incluir el proceso NER como parte del sistema TBIR.

7.1.3.1 Con colección de Wikipedia en *ImageCLEF* 2011

Analizando la parte textual del conjunto de consultas propuestas en la edición 2011 del *ImageCLEF* (Tabla 4-5), se observa que la mayoría de ellas no incluyen entidad nombrada alguna: solo en 8 de las 50 consultas se reconoce alguna entidad (haciendo uso de la

herramienta *Stilus*). Por este motivo, y para poder extraer alguna conclusión acerca de la aportación del proceso NER, se lleva a cabo una serie de experimentos únicamente sobre este subconjunto de consultas que contienen entidades. Estos resultados fueron publicados como parte de la experimentación presentada en la correspondiente edición del *ImageCLEF* (Granados et al., 2011). Se describen a continuación los experimentos desarrollados para su posterior comparación.

- baseline. Experimento textual básico en el que se aplica la mejor configuración del preprocesamiento analizado en el apartado 7.1.1 (stopwords, stemming, etc.) y se utiliza información textual adicional extraída de Wikipedia (apartado 7.1.2). No hay multilingüismo, se hace uso únicamente de la información disponible para el idioma inglés.

- *earlyEnrich*. Sigue la misma configuración que el experimento anterior (*baseline*), pero añade las entidades nombradas (NE) identificadas en el texto asociado a cada imagen como parte del mismo, esto es, enriquece los metadatos correspondientes a cada imagen con las entidades. Los términos que forman las entidades detectadas no son procesados lingüísticamente, sino que mantienen su forma original (sin eliminar *stopwords*, sin aplicar *stemming*, etc.). Este mismo proceso de enriquecimiento es realizado con la parte textual de las consultas multimedia, las cuales serán lanzadas contra el índice generado a partir de las anotaciones enriquecidas de las imágenes.

- *lateEnrich*. En este caso se generan dos índices independientes: uno igual que en el experimento textual base (*baseline*), y otro haciendo uso únicamente de las entidades nombradas. De igual modo, se generarán dos tipos de consultas: las que contienen el todo el texto de la misma (como en *baseline*), y las que solo están formadas por las entidades identificadas. Las dos listas de resultados obtenidas serán combinadas siguiendo una aproximación de fusión tardía (*late fusion*) que generará la lista de resultados final. El algoritmo de fusión utilizado será el *Enrich* (descrito en el apartado 6.4), utilizando como lista principal la obtenida por el *baseline*, y como secundaria la resultante de utilizar únicamente entidades.

Se muestran en la siguiente tabla los resultados obtenidos para cada uno de estos experimentos, incluyendo los valores de precisión media (AP) sobre las 8 consultas con

entidades, y los correspondientes a la precisión alcanzada en las 5 primeras imágenes devueltas (P@5).

Tabla 7-6. Resultados en consultas con entidades nombradas

Consulta	*baseline*		*lateEnrich*		*earlyEnrich*	
	AP	**P@5**	**AP**	**P@5**	**AP**	**P@5**
71	0,4754	0,6000	0,4190	0,6000	0,2256	0,6000
84	0,6552	0,6000	0,6193	0,6000	0,4210	0,4000
85	0,3969	0,6000	0,1803	0,2000	0,1194	0,4000
88	0,7886	0,8000	0,7734	0,8000	0,6989	0,8000
89	0,5515	0,4000	0,4699	0,4000	0,5547	0,4000
90	0,1215	0,6000	0,0845	0,2000	0,0965	0,2000
113	0,3435	1,0000	0,3582	1,0000	0,3708	1,0000
118	0,0352	0,0000	0,0349	0,0000	0,0458	0,2000
Media	0,4210	0,5750	0,3674	0,4750	0,3166	0,5000

Analizando la tabla, puede observarse que los resultados (tanto en términos de MAP como de P@5) no son significativamente mejorados por los experimentos que hacen uso de la información obtenida mediante el proceso de NER. Solo en casos aislados para algunas consultas (89, 113 y 118), y de manera muy leve, se mejoran los valores obtenidos por el experimento *baseline*. En cuanto a las dos estrategias de enriquecimiento en base a las entidades, funciona mejor la aproximación *lateEnrich* que combina a nivel de resultados (*late fusion*) las listas de imágenes obtenidas ($MAP = 0.3674$), en comparación con *earlyEnrich* que concatena toda la información textual con la correspondiente a las entidades desde un primer momento ($MAP = 0.3166$).

Para profundizar algo más y tratar de entender los motivos que hacen que el enriquecimiento llevado a cabo en base a la información sobre las entidades nombradas no solo no mejore los resultados originales, sino que en algunos casos los empeore, se analizan ejemplos los concretos observados. De las 8 consultas que contienen imágenes, la primera de ellas (71) se corresponde con el texto "*coloured Wolkswagwen beetles*", en la cual se identifican las entidades "*Wolkswagen Group*" y "*Wolkswagen*". Para esta consulta en concreto ninguno de los enriquecimientos analizados mejora los valores de MAP ni de P@5 obtenidos por el *baseline*. Parece ser que la información sobre las entidades añadida no aporta positivamente e incluso parece introducir ruido a tenor de los resultados obtenidos, ya que la herramienta de

NER utilizada (*Stilus*) proporciona en algunos casos información extra sobre la entidad identificada. Otro ejemplo de posible introducción de ruido es el caso de la consulta 84 ("*Sagrada Familia in Barcelona*"), para la que la herramienta devuelve las entidades "*Basílica i Temple Expiatori de la Sagrada Família*", "*Sagrada Familia*" y "Barcelona". Otro tipo de consultas son aquellas en las que el NER no introduce términos textuales nuevos (85, 88, 89, 90 y 113), sino que simplemente identifica palabras de la misma que son entidades nombradas. Por ejemplo, en la consulta 85 ("*Beijing bird nest*") simplemente se identifica "*Beijing*" como entidad nombrada, por lo que se repetirá en la consulta enriquecida de la aproximación *earlyEnrich*, o simplemente se usará como consulta en el caso del *lateEnrich*. Por último, resaltar el caso de la consulta 118 ("*flag of UK*"), en la que la identificación de la entidad "*UK*" y de la información extra "*United Kingdom*" proporcionada por le herramienta *Stilus*, resulta beneficiosa para la mejora de los resultados de evaluación: el valor de MAP se ve incrementado ligeramente siguiendo la aproximación de *earlyEnrich*, así como el valor de P@5, que pasa de 0,0 a 0,2 (lo que significa que si antes no se recuperaba ninguna imagen relevante entre las 5 primeras devueltas, ahora si aparecería una). En resumen, la información adicional obtenida gracias al NER puede resultar útil en algunos casos, siempre y cuando esta información se integre dentro del proceso de recuperación de la manera correcta, evitando la introducción de ruido que empeorará los resultados. En la experimentación aquí mostrada, los resultados no mejoran en la mayoría de los casos y, además, se identifican entidades en pocas de las consultas propuestas para la evaluación de la tarea de recuperación.

7.1.3.2 Con colección IAPR TC-12

Los experimentos mostrados en este apartado (Agerri, Granados and Garcia-Serrano, 2011) son realizados sobre la colección IAPR TC-12, proporcionada en la edición de 2008 de la tarea de recuperación de imágenes fotográficas del *ImageCLEF*. Las consultas multimedia proporcionadas por los organizadores son un total de 39, de las que 19 de ellas tienen alguna entidad nombrada en su parte textual.

En el tratamiento de las entidades nombradas llevado a cabo en esta evaluación, se distingue entre entidades nombradas comunes (NEs, entendidas como asignaciones rígidas ya que se refieren unívocamente a una entidad en el mundo) y asignaciones no rígidas (NRDEs, *non-rigid designators*, que simplemente se refieren a una entidad de una manera no única). Ejemplos de NEs podrían ser "España" o "Los Alpes", mientras que para NRDEs se tendrían

entidades como “el presidente de la Comisión Europea” o “el corresponsal de El País”. Se quiere analizar con los experimentos planteados a continuación la influencia de negativa que puede tener el no extraer estos modificadores asociados a ciertas entidades nombradas.

Para el proceso de identificación de entidades se utiliza un *pipeline* de etiquetadores de C&C[27]: tokenizador, POS (*part-of-speech*), y NER. Se mantienen únicamente las entidades anotadas como lugar, persona y organización, ya que el resto de categorías (expresiones numéricas como fechas, horas, porcentajes y dinero) no son tenidas en cuanta durante el proceso de indexación.

Se llevan a cabo diferentes experimentos para evaluar la influencia de este proceso durante el proceso de anotación y recuperación de imágenes:

- *baseline*. Experimento basado en una recuperación textual que construye el índice con los metadatos incluidos en los campos <*title*>, <*description*>, <*notes*> y <*location*>. Para construir las consultas textuales correspondientes se hace uso de los campos de información <*title*>, <*topic*> y <*narration*>, incluidos en la descripción de cada consulta. El proceso de indexación y recuperación en llevado a cabo mediante la herramienta IDRA.

- NER. El proceso de indexación y recuperación es igual que en el experimento anterior, pero en este caso se añaden los campos <*ner*> y <*nrde*> al índice. Estos campos contienen respectivamente las entidades nombradas y las entidades nombradas con modificadores (NRDEs) detectadas en la información textual asociada a cada imagen. La construcción de las consultas añade, en comparación con el *baseline*, las entidades nombradas detectadas en el texto de las consultas.

- NRDE. Este experimento utiliza el mismo índice que el anterior (con los campos <ner> y <nrde> correspondientes a las entidades y a su versión modificada). Las consultas añadirán a las utilizadas en el *baseline*, las NRDEs identificadas.

Dado el bajo número de entidades nombradas presentes tanto en los metadatos de las imágenes como en las consultas, la primera conclusión que podría extraerse es que el proceso

[27] http://svn.ask.it.usyd.edu.au/trac/candc

NER no es recomendable para la tarea descrita sobre la colección propuesta. En cualquier caso, se puede comparar el rendimiento obtenido para cada una de las 39 consultas en función de si éstas contienen o no alguna entidad nombrada. Como se verá, los resultados globales de los experimentos con entidades (enriquecidas o no) son muy similares a los obtenidos por el *baseline*.

Se muestra, para cada consulta, la precisión global obtenida (MAP), las precisiones tempranas a 5 y a 20 (P@5 y P@20), si en ella se identificó alguna entidad nombrada estándar o enriquecida (NER/NRDE).

Tabla 7-7. Impacto de NER/NRDE

Q	*baseline*			NER			NRDE			NER/ NRDE
	P@5	P@20	MAP	P@5	P@20	MAP	P@5	P@20	MAP	
2	0,0000	0,0000	0,0280				0,0000	0,0000	0,0291	-
3	0,2000	0,2500	0,2368				0,0000	0,3000	0,2161	-
5	0,0000	0,1000	0,0940				0,2000	0,2500	0,1277	-
6	0,2000	0,2500	0,2388	0,6000	0,6500	0,3983		0,6500	0,3983	✓
10	0,6000	0,6000	0,5034	0,6000	0,6500	0,4751	0,6000	0,6500	0,4760	✓
11	0,6000	0,3000	0,3021	0,6000	0,4500	0,4200	0,6000	0,4500	0,4200	✓
12	0,4000	0,2000	0,1534				0,2000	0,3500	0,1562	-
13	0,6000	0,3500	0,2251				0,4000	0,3500	0,1831	-
15	0,0000	0,0000	0,0443				0,0000	0,0500	0,0441	-
16	0,0000	0,1500	0,1735	0,0000	0,0500	0,2446	0,0000	0,0500	0,2446	✓
17	0,8000	0,6000	0,5593				1,0000	0,6000	0,6371	-
18	0,0000	0,3000	0,1587	0,0000	0,0000	0,0267	0,0000	0,0000	0,0267	✓
19	0,2000	0,1500	0,0613				0,0000	0,1500	0,0538	-
20	0,2000	0,0500	0,0079				0,2000	0,0500	0,0098	-
21	0,0000	0,0500	0,2511				0,2000	0,2000	0,2788	-
23	0,0000	0,5000	0,1343	1,0000	0,6500	0,2194	1,0000	0,6500	0,2194	✓
24	0,0000	0,0500	0,0197	0,0000	0,1000	0,0190	0,0000	0,1000	0,0190	✓
28	0,2000	0,2000	0,0899	0,2000	0,1000	0,0479	0,2000	0,1000	0,0479	✓
29	0,8000	0,9000	0,7720	0,6000	0,4000	0,4067	0,6000	0,4500	0,5146	✓
31	0,8000	0,5000	0,3272	0,4000	0,2000	0,1438	0,4000	0,2000	0,1438	✓
34	0,8000	0,6000	0,3077	0,4000	0,2500	0,1154	0,2000	0,1000	0,1395	✓
35	0,8000	0,8500	0,6108				0,8000	0,9000	0,6166	-
37	0,2000	0,1500	0,0796	0,0691	0,0000	0,0691	0,0000	0,0000	0,0691	✓
39	0,0000	0,2667	0,2308				0,0000	0,3000	0,2401	-
40	0,0000	0,1500	0,0779				0,2000	0,1000	0,0748	-
41	0,0000	0,0000	0,0204	0,0000	0,0500	0,0200	0,0000	0,0500	0,0200	✓

43	0,0000	0,2500	0,1797				0,2000	0,2000	0,1824	-
44	0,0000	0,1000	0,0437	0,0000	0,3000	0,0442	0,0000	0,1000	0,0480	✓
48	0,0000	0,2000	0,1627	0,0000	0,0000	0,0933	0,0000	0,0000	0,0933	✓
49	0,8000	0,5000	0,1505	0,8000	0,6000	0,1609	0,8000	0,6000	0,1609	✓
50	0,6000	0,2500	0,1497				0,4000	0,2000	0,1471	-
52	0,6000	0,3000	0,2459				0,6000	0,3500	0,2617	-
53	0,8000	0,5500	0,3945				0,6000	0,4500	0,3349	-
54	0,8000	0,5500	0,6464	0,8000	0,4500	0,5349	1,0000	0,5000	0,5968	✓
55	0,2000	0,0500	0,0126	0,2000	0,0500	0,0126	0,2000	0,0500	0,0126	✓
56	0,0000	0,2000	0,2240				0,0000	0,1500	0,2053	-
58	0,8000	0,4500	0,3599				0,8000	0,4000	0,3242	-
59	0,2000	0,1500	0,1324	0,0000	0,0500	0,0873	0,0000	0,0500	0,0873	✓
60	0,8000	0,7500	0,6268				1,0000	0,7500	0,7166	-

Analizando los resultados mostrados en la tabla anterior se distinguen 3 tipos de comportamiento. En primer lugar se observan varios casos en los que el tratamiento NER+NRDE mejora el *baseline* (consultas 6, 10, 11, 23 y 49). Dado que el sistema es capaz de extraer entidades nombradas enriquecidas (NRDE) del texto de estas consultas, es posible otorgar una ponderación extra a dichos términos con lo que se consigue una precisión mayor que con el experimento *baseline* (por ejemplo la parte textual de la consulta 23 es "*sport activities in the US state of California*", de la que se pueden extraer las entidades "*California*" y "*US state of California*").

Otro tipo de comportamiento (observado en las consultas 16, 18, 24, 28, 37, 41, 44, 48 y 54) muestra cómo el proceso de extracción de entidades no influye en los resultados obtenidos, donde se ve que los valores de evaluación son muy similares a los correspondientes al *baseline*. Esto indica que el tratamiento dado a estas consultas debería ser más minucioso. Por ejemplo, el tratamiento dado en la consulta 16 ("*images of San Francisco with at least one person*") a las entidades reconocidas ("*San Francisco*") no supone ninguna mejora de rendimiento en relación a la obtenida por el *baseline*. En ninguno de los dos se consigue identificar la no relevancia de las imágenes de San Francisco sin ninguna persona. El mismo fenómeno se puede ver en la consulta 18, en la que se buscan estadios deportivos fuera de Australia ("*sport stadium outside Australia*") y el etiquetador de entidades identifica simplemente "*Australia*". Un análisis más profundo del texto de las consultas (como la

identificación de la negación) podría resultar beneficioso para mejorar el rendimiento del sistema de recuperación.

Por último, en algunos casos (consultas 29, 31, 34 y 54), se observa que el tratamiento de entidades empeora los resultados del *baseline*. La explicación es similar a la del caso anterior, solo que ahora los resultados son peores debido al gran número de entidades erróneamente detectadas. Un ejemplo puede verse en la consulta 31 ("*volcanoes around Quito*"), que proporciona una larga lista de nombres de volcanes. El sistema identifica todos estos nombres como entidades, pero no coge el término "volcán", y dado que dichos nombres también corresponden a ciudades de Ecuador, se introduce mucho ruido en el proceso de recuperación. Esto hace que se recuperan muchas imágenes no relevantes, lo que empeora los valores de evaluación obtenidos. Las consultas 29 ("*views of Sydney's world-famous landmarks*") y 34 ("*group picture on a beach*") contienen ejemplos similares.

Finalmente, destacar que, aunque no se detectan demasiadas entidades nombradas enriquecidas ni en las consultas ni en las anotaciones de las imágenes, el experimento NRDE mejora el rendimiento del NE. Esto implica que el procedimiento de enriquecimiento de las entidades parece tener un impacto positivo en los resultados finales.

7.1.4 Recuperación multilingüe

La información textual asociada a las imágenes, o a cualquier objeto multimedia, puede estar disponible en varios idiomas, como en el caso de la colección de imágenes de Wikipedia (apartado 4.2.1.3). Aparece entonces la necesidad de combinar esta información con el objetivo de dilucidar por un lado si de dicha colaboración surgirá un sistema TBIR multilingüe con mejor rendimiento y, por otro, la mejor manera de llevar a cabo dicha combinación o fusión textual.

Para evaluar la aportación de la combinación de la aproximación multilingüe, se comparará esta con los resultados de la recuperación textual basada únicamente en inglés, que es el idioma con el que mejores resultados se han obtenido (como puede verse en el apartado 7.1.2).

Para llevar a cabo la fusión textual existen dos aproximaciones clásicas, en función del nivel al que se realice dicha fusión (ver apartado 3.2): nivel de características o fusión temprana

(*early fusion*) y nivel de decisiones o fusión tardía (*late fusion*). Las características en el caso de trabajar con texto serán las palabras o términos de indexación, por lo que la fusión temprana consistirá en concatenar los términos procedentes de cada idioma, esto es, juntar todo el texto independientemente del idioma del que provenga, previamente a la fase de indexación/recuperación. Por otro lado, la solución basada en fusión tardía consistirá en completar la recuperación textual monolingüe para cada uno de los idiomas y, posteriormente, fusionar las listas de resultados obtenidas desde cada uno de ellos. La técnica de fusión tardía empleada en estos experimentos es el algoritmo *combMAX*, descrito detalladamente en el apartado 3.4, que seleccionará para cada consulta las imágenes recuperadas con mayor valor de relevancia (independientemente del idioma).

En la tabla siguiente pueden verse los resultados tanto de la recuperación monolingüe en inglés, como los obtenidos tras los dos tipos de fusión textual descritos. En la tabla "*early*" indica fusión temprana, y "*late*" fusión tardía.

Tabla 7-8. Fusión textual multilingüe

	Idioma	**Fusión textual**	**MAP**	**Mejora**
Experimento1	ingles	-	0,2527	-
Experimento2*	todos	early	0,2859	13,14%
Experimento3*	todos	late	0,2885	14,17%

(*) Mejoras estadísticamente significantes

Los resultados obtenidos indican que la información textual disponible en alemán y francés aporta nuevo conocimiento a la existente en inglés, esto es, son informaciones complementarias ya que mejoran los resultados monolingües en inglés en más de un 13% y un 14% respectivamente, en función de la estrategia de fusión seguida. De entre los dos tipos de fusión planteados (*early fusion* y *late fusion*), los resultados son ligeramente mejores en el segundo caso, esto es, cuando primero se recuperan imágenes en base a cada idioma y, posteriormente, se fusionan esos resultados utilizando *combMAX*.

7.1.5 Configuración óptima para TBIR

Una vez analizadas y evaluadas (sobre el conjunto de *topics* de la edición de 2010 de la tarea de recuperación de imágenes de Wikipedia de *ImageCLEF*) las distintas posibilidades en cuanto a la configuración de los diferentes parámetros relacionados con el subsistema textual

de recuperación de imágenes (TBIR), se diseña la versión final de éste teniendo en cuenta las soluciones que mejores resultados han proporcionado:

- Lista de *stopwords*. Se utilizará la lista de palabras de parada disponible en la herramienta IDRA, que funciona mejor para la tarea de recuperación de imágenes que la facilitada por defecto en Lucene. La eliminación de las palabras de esta lista se llevará a cabo tanto en el texto asociado a las imágenes de colección, como en el de las consultas de evaluación.

- *Stemming*. Se incluye esta técnica en el preprocesamiento textual de las anotaciones asociadas a las imágenes, previamente a su indexación. Igualmente se aplicará al preprocesar la parte textual de las consultas multimedia.

- Metadatos. La información textual del campo <*comment*>, independiente del idioma, proporcionada como parte de los metadatos, será tenida en cuenta, preprocesada e indexada por el subsistema TBIR.

- Herramienta de indexación /recuperación. Los resultados obtenidos con la herramienta IDRA y con Lucene son bastante parecidos. La decisión final consiste en utilizar una combinación de ambas: se empleará IDRA para la fase de preprocesamiento textual, y Lucene para las tareas de indexación y recuperación, en las que es bastante más eficiente en cuanto al tiempo de respuesta.

- Enriquecimiento textual basado en Wikipedia. Se utilizará la información textual disponible en los campos <*title*> y <*categories*> de los artículos de Wikipedia en la que aparecen las imágenes de la colección.

- Reconocimiento de entidades nombradas (NER). En los experimentos observados este proceso no mejora globalmente los resultados de la recuperación, por lo que no se incluirá en la configuración TBIR final.

- Fusión textual multilingüe. Se ha comprobado que utilizar la información textual disponible en todos los idiomas genera mejores resultados que emplear únicamente la del mejor idioma, en este caso el inglés. Además, la fusión tardía (*late fusion*) de

las anotaciones textuales de cada idioma funciona mejor que la fusión temprana (*early fusion*).

En base a todas estas decisiones se diseña e implementa el sistema final de recuperación de imágenes basado en texto (TBIR). En el siguiente apartado, esta configuración será evaluada sobre la misma colección de imágenes de Wikipedia pero, en este caso, sobre el conjunto de consultas de evaluación o *topics* proporcionados en el *ImageCLEF* 2011 con el objetivo de corroborar las decisiones tomadas, ya que todos los experimentos anteriores (a excepción de los de las entidades nombradas) fueron evaluados en base a las consultas facilitadas en *ImageCLEF* 2010.

7.1.6 Consolidación de la configuración TBIR

Se vuelven a evaluar, sobre el conjunto de consultas de *ImageCLEF* 2011, los siguientes parámetros: la conveniencia de incluir el stemming en la fase de preprocesamiento, la utilidad de tener en cuenta la información textual extraída de los artículos de Wikipedia, el comportamiento de las herramientas IDRA y Lucene para la fase de indexación/recuperación y, por último, la estrategia de fusión a seguir para combinar la información de cada idioma.

Se muestran en primer lugar un conjunto de experimentos relacionados con la aplicación de *stemming*, con el uso de la información extraída desde Wikipedia, y con la conveniencia de seguir una aproximación multilingüe. La siguiente tabla muestra los resultados obtenidos, indicando la mejora obtenida en cada caso en relación al primer experimento (en el que no se aplica stemming, ni enriquecimiento Wikipedia, ni multilingüismo):

Tabla 7-9. Stemming, enriquecimiento Wikipedia, multilingüismo

	Idioma	***Stemming***	**Enriquecimiento Wikipedia**	**MAP**	**Mejora**
Experimento1	ingles	NO	NO	0,1727	-
Experimento2		SI		0,2056	19,05%
Experimento3			SI	0,2243	29,88%
Experimento4	todos			0,2489	44,12%

Puede observarse cómo las decisiones tomadas a lo largo de este capítulo, y resumidas en el apartado 7.1.5, fueron acertadas y se confirman con este nuevo conjunto de experimentos (con otras consultas).

La siguiente tabla corresponde a la comparación del rendimiento de indexación y recuperación de las herramientas IDRA y Lucene:

Tabla 7-10. Herramienta de indexación / recuperación

	Sistema	**Lang**	**MAP**	**Mejora**
Experimento1	IDRA	EN	0,2243	15,96%
	Lucene		0,2601	
Experimento1	IDRA	ALL	0,2489	22,30%
	Lucene		0,3044	

Para este conjunto de consultas la herramienta Lucene proporciona un mejor rendimiento. Se confirma entonces la decisión tomada de llevar a cabo el preprocesamiento textual utilizando la herramienta IDRA, y la parte de indexación y recuperación con Lucene.

Por último, el aspecto referido a la forma de combinar la información textual procedente de los distintos idiomas existentes en la colección es evaluado comparando la técnica de fusión a nivel de características (*early fusion*) con la de fusión a nivel de decisiones (*late fusion*). Los resultados obtenidos se muestran en la siguiente tabla:

Tabla 7-11. Estrategia para recuperación multilingüe

Sistema	**Stemming**	**Wikipedia**	**Fusión textual**	**MAP**	**Mejora**
Lucene	Si	Si	early	0,2758	10.37%
			late	0,3044	

También en este caso la decisión tomada para configurar el subsistema TBIR se confirma, ya que la estrategia de fusión de idiomas que mejores resultados obtiene es la basada en las decisiones individuales tomadas en base a la información textual de cada idioma, esto es, *late fusion*.

7.1.7 Comparación con otras aproximaciones TBIR

Evaluando globalmente el conjunto total de *topics* propuestos en las dos ediciones de la tarea de recuperación de imágenes de Wikipedia del *ImageCLEF* (2010 y 2011), los resultados obtenidos por el sistema TBIR definido en esta tesis (apartado 6.2) son los mejores tanto en términos de precisión global (MAP) como de precisiones bajas (P@10 y P@20), si se compara con el resto de aproximaciones presentadas a la competición (Tabla **7-14**). Analizando los resultados independientemente para cada edición, la propuesta TBIR presentada hubiera ganado en 2010 (Tabla **7-12**) y quedado en segunda posición en 2011 (Tabla **7-13**).

La tabla siguiente muestra los resultados obtenidos por la configuración TBIR propuesta en esta tesis, en comparación con las mejores aproximaciones textuales del resto de grupos presentadas en la edición de *ImageCLEF* 2010. Se muestran los resultados de evaluación obtenidos para las medidas P10, P20 y MAP, y se incluye la media calculada sobre todos los grupos participantes (teniendo en cuenta únicamente los resultados pertenecientes al 90% superior, para evitar resultados ruidosos o erróneos):

Tabla 7-12. Comparación TBIR en *ImageCLEF 2010*

Grupo	MAP	P@10	P@20
TBIR	**0,2885**	**0,5414**	**0,4971**
xrce	0,2361	0,4871	0,4393
unt	0,2251	0,4314	0,3871
telecom	0,2227	0,4829	0,4407
i2rcviu	0,2126	0,4486	0,4143
dcu	0,2039	0,4271	0,3907
cheshire	0,2014	0,4600	0,4036
daedalus	0,1820	0,4471	0,4029
duth	0,1818	0,4243	0,4079
sztaki	0,1768	0,4714	0,4229
nus	0,1581	0,3529	0,3264
MEDIA	0,1602	0,4024	0,3575

En total fueron presentados 48 experimentos textuales por 11 grupos de investigación diferentes, de los cuales 18 de ellos utilizan técnicas de expansión de consultas (*query expansion*) y 23 siguen aproximaciones basadas en realimentación por relevancia (*relevance*

feedback) o pseudo relevancia (Popescu, Tsikrika and Kludas, 2010). Por el contrario, la aproximación textual seguida en esta tesis no utiliza ninguno de esos mecanismos, tal y como se ha ido describiendo a lo largo de esta sección.

El grupo que obtiene los segundos mejores resultados (*xrce*, *Xerox Research Centre Europe*) utiliza una representación de los metadatos textuales basada en modelos del lenguaje estándar (Clinchant et al., 2010). Este grupo, al igual que la configuración TBIR planteada en esta tesis, también utiliza la información textual de los artículos de Wikipedia en los que aparecen las imágenes, extrayendo de ellos el párrafo en el que esta es mencionada. Otros grupos (*sztaki* y *dcu*) utilizan Okapi BM25 como función de recuperación textual (Daroczy, Petras and Benczur, 2010) (Min, Laveling and Jones, 2010). Este último grupo (dcu) además lleva a cabo una expansión de documentos en base al contenido de Wikipedia. El grupo que obtiene el peor valor de MAP para sus experimentos textuales (*nus*) también utiliza la información de Wikipedia, mapeando los metadatos de cada imagen sobre conceptos de la enciclopedia. El grupo *telecom* utiliza también la información proporcionada por Wikipedia, pero en este caso para expandir las consultas en base a conceptos relacionados (Popescu, 2010). Otras técnicas de expansión de consulta y de corpus son aplicadas por el grupo *daedalus* (Lana-Serrano, Villena-Román and González-Cristobal, 2010), así como información sobre entidades nombradas y conceptos incluidos en *DBpedia*[28]. Por último, comentar la estrategia manual para la expansión de la consulta seguida por el grupo *unt* (Ruiz et al., 2010), con el objetivo de añadir palabras útiles para ayudar a completar la consulta. Esta técnica de recuperación interactiva mejora la precisión de los primeros resultados, pero no el rendimiento global del sistema.

En relación a la edición de *ImageCLEF* 2011, se incluye la Tabla **7-13** que igualmente muestra los mejores resultados obtenidos por cada uno de los grupos participantes en la modalidad de recuperación de imágenes basada en texto. Se incluyen también los valores medios de evaluación teniendo en cuenta el 90% mejor de todos los experimentos textuales presentados.

[28] http://es.dbpedia.org/

Tabla 7-13. Comparación TBIR en *ImageCLEF 2011*

Grupo	FB/QE	MAP	P@10	P@20
xrce	FBQE	0,3141	0,5160	0,4270
TBIR	NOFB	0,3044	0,5060	0,4040
untesu	FB	0,2866	0,4220	0,3650
cea list	QE	0,2591	0,4660	0,3630
demir	NOFB	0,2369	0,4180	0,3320
redcad	NOFB	0,2306	0,3700	0,3060
sztaki	NOFB	0,2136	0,4380	0,3480
sinai	FB	0,2068	0,4020	0,3380
dbisformat	NOFB	0,2043	0,3960	0,3120
MEDIA	-	0,2189	0,4007	0,3254

Para esta última edición de la tarea de recuperación de imágenes de Wikipedia se presentaron 51 experimentos textuales, correspondientes a 9 grupos de investigación diferentes. Como puede observarse en la tabla, el rendimiento obtenido por la aproximación textual (TBIR) propuesta en esta tesis supera a todas las restantes, exceptuando la del grupo de Xerox cuyos resultados son ligeramente mejores. Dicha aproximación textual ganadora implementa un modelo de recuperación de información basado en implicación léxica (Csurka, Clinchant and Popescu, 2011). Del resto de grupos participantes, 16 de los experimentos presentados aplican expansión de la consulta, y 15 siguen técnicas de *relevance feedback* (Tsikrika, Popescu and Kludas, 2011).

Los resultados obtenidos en base a la aproximación TBIR propuesta no utilizan ni expansión de consulta ni realimentación por relevancia, cosa que si hacen la mayoría de experimentos presentados. Por ejemplo, el tercer grupo clasificado (*untesu*) aplica análisis semántico saliente (*Salient Semantic Analysis*) para expandir la consulta con términos semánticamente similares de Wikipedia (Ruiz, Leong and Hassan, 2011). También el grupo *cea list* utiliza Wikipedia para expandir la consulta con términos relacionados (Csurka, Clinchant and Popescu, 2011). Otra alternativa para expandir es la seguida por los grupos *sztaki* (Daroczy, Pethes and Benczur, 2011) y *uaic* (Boros, Ginsca and Iftene, 2011), que añaden sinónimos obtenidos desde *WordNet* a los términos originales de la consulta textual. En cuanto a las herramientas de indexación y recuperación textual utilizadas suele utilizarse *Lucene*, como en el caso de los grupos *uaic* o *redcad* (Awadi, Khemakhem and Jemaa, 2011), pero también

otras como la plataforma *Terrier*[29], utilizada por el grupo *demir* (Berber et al., 2011) para evaluar el rendimiento de distintos esquemas clásicos de pesado (BM25, *tf-idf*, etc.).

Para terminar la comparación de la aproximación TBIR propuesta en esta tesis con el resto de soluciones observadas, se recopilan en la Tabla 7-14 los resultados obtenidos junto con la mejor aproximación textual presentada tanto en la edición de 2010 como en la de 2011, así como los resultados que se obtendrían sobre el conjunto global de consultas para ambas ediciones (un total de 120 consultas). Se incluye también el valor medio de los resultados obtenidos por todos los experimentos presentados por los diferentes grupos que participaron en ambas ediciones.

Tabla 7-14. Comparativa TBIR en *ImageCLEF* (2010+2011)

	2010+2011		
	MAP	**P@10**	**P@20**
TBIR	**0,2951**	**0,5267**	**0,4583**
mejor	0,2686	0,4991	0,4342
MEDIA	0,1825	0,3966	0,3398

Sobre el conjunto global de consultas obtenido de unir las proporcionadas para ambas ediciones, el rendimiento de TBIR alcanza un 29,51% en términos de MAP, superior al que obtendría Xerox (26,86%). También los valores de evaluación obtenidos para las medidas de precisión en los primeros resultados (P@10 y P@20) son mejorados por la aproximación TBIR. El rendimiento medio de los sistemas de recuperación textual presentados durante las ediciones 2010 y 2011 queda considerablemente lejos de la propuesta en esta tesis, tanto en términos de MAP (18,25%) como de precisiones bajas ($P@10 = 39{,}66\%$, $P@20 = 33{,}98\%$).

El hecho de obtener unos resultados textuales de recuperación de imágenes competitivos es importante dentro del trabajo presentado en esta tesis, ya que la aproximación de fusión multimedia que finalmente se propone parte de una fase de prefiltrado semántico textual que se apoya inicialmente en el sistema TBIR. Como se verá en el apartado 7.3.1, dedicado a los experimentos relacionados con el prefiltro semántico textual (STP), no solo los valores de

[29] http://terrier.org/

precisión del subsistema TBIR son positivos, sino que también la cobertura de los resultados es bastante alta (casi un 83%), lo que también resultará ser un dato importante sobre el que se sostendrá la estrategia final de recuperación de imágenes en base a la técnica de fusión multimedia semántica tardía (LSMF) propuesta.

7.2 Recuperación de imágenes basada en contenido (CBIR)

La experimentación en cuanto a la recuperación de imágenes basada en su contenido visual (CBIR) ha sido desarrollada en colaboración con investigadores de la Universidad de Valencia (UV), y han permitido analizar las medidas de similitud empleadas para calcular la semejanza entre las imágenes de la colección y las proporcionadas como ejemplo en una consulta multimedia. También se evalúan diferentes estrategias de agregación para combinar los valores de semejanza obtenidos entre las imágenes de la colección y cada una de las proporcionadas como parte de la consulta multimedia, y se comparan diferentes algoritmos de recuperación (automático, expansión de la consulta, y regresión logística). Finalmente, se incluye un grupo de experimentos dedicados a evaluar cómo funcionan determinados grupos de descriptores visuales trabajando de manera individual en la tarea de recuperación.

7.2.1 Configuración del sistema CBIR

Un **primer conjunto de experimentos** (Garcia-Serrano et al., 2009), presentados en la tarea de recuperación de imágenes fotográficas sobre la colección IAPR TC-12 (descrita en 4.2.1.1), tiene como objetivo evaluar el rendimiento del sistema CBIR en función de las distintas configuraciones de dos parámetros:

1) la elección del tipo de distancia para calcular la similitud entre los vectores de características de las imágenes de la colección y las proporcionadas como ejemplo dentro de cada *topic* o consulta multimedia. Se trabaja con 3 distancias distintas:

    - distancia euclídea. En un espacio de dos dimensiones, la distancia euclídea entre dos puntos X e Y de coordenadas (x_1, x_2) y (y_1, y_2) respectivamente, se calcularía en base a la siguiente fórmula:

$$d_E(X,Y) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- distancia de *Mahalanobis*. Se diferencia de la euclídea en que tiene en cuenta la correlación entre las variables aleatorias entre las que trata de determinar su similitud. Formalmente, la distancia de *Mahalanobis* entre dos variables aleatorias con la misma distribución de probabilidad x e y con matriz de covarianza Σ se define como:

$$d_m(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$$

2) la estrategia de agregación para combinar las semejanzas (distancias de similitud) obtenidas por las imágenes de la colección con respecto a cada una de las imágenes que se proporcionan como ejemplo dentro de cada consulta multimedia. Se experimenta con cinco operadores de agregación OWA: máximo (OR), media, mínimo (AND), $orness(W) = 0.3$ y $orness(W) = 0.7$.

Los resultados que se obtienen tras el proceso de recuperación de imágenes basado únicamente en las características visuales de las mismas, y combinando diferentes configuraciones para los parámetros mencionados, son los que se muestran en la siguiente tabla:

Tabla 7-15. Sistema CBIR: distancia y operador OWA

	Distancia	**OWA**	**MAP**	**P@5**	**P@10**	**P@20**
Experimento1	Euclídea	máximo	0,0042	0,0256	0,0128	0,0103
Experimento2		media	0,0073	0,0667	0,0487	0,0282
Experimento3		mínimo	0,0137	0,1179	0,0667	0,0487
Experimento4		owa(0,3)	0,0110	0,0923	0,0615	0,0359
Experimento5		owa(0,7)	0,0033	0,0154	0,0077	0,0077
Experimento6	Mahalanobis	máximo	0,0050	0,0410	0,0256	0,0244
Experimento7		media	0,0067	0,0359	0,0333	0,0308
Experimento8		mínimo	**0,0213**	**0,1744**	**0,1026**	**0,0679**
Experimento9		owa(0,3)	0,0105	0,0615	0,0462	0,0385
Experimento10		owa(0,7)	0,0057	0,0359	0,0256	0,0269

De los 10 experimentos realizados, el que mejores resultados obtiene para todas las medidas de evaluación mostradas es el Experimento8, cuya configuración está basada en el uso de la distancia de *Mahalanobis* y del operador de agregación OWA mínimo (consistente en seleccionar el valor más bajo de entre todos los de entrada).

Analizando con un poco más de detalle toda la tabla de resultados, puede observarse que para todos los casos, la distancia de *Mahalanobis* ofrece un mejor rendimiento. Lo mismo sucede para la estrategia de agregación, para la que el OWA mínimo (AND) resulta la mejor opción, seguido del $orness(W) = 0.3$ que es un AND suavizado.

Un segundo conjunto de experimentos (Benavent et al., 2010), esta vez sobre la colección de imágenes de Wikipedia y utilizando las consultas del *ImageCLEF* 2010 (*topics2010*), se centra en evaluar tres algoritmos de recuperación diferentes para implementar el subsistema CBIR, y en cómo estos asignarán valores de relevancia (*scores* visuales, S_v) a cada una de las imágenes de la colección para cada consulta multimedia concreta. Los tres algoritmos implementados y evaluados son los siguientes:

- Algoritmo automático. Se trata del clásico algoritmo en sistemas CBIR, que se basa en el cálculo del vector de características de bajo nivel que describe cada imagen de la colección. A continuación calcula las medidas de similitud entre el vector de características de cada imagen y el de cada una de las imágenes de ejemplo proporcionadas como parte de las consultas multimedia. La distancia utilizada es la de *Mahalanobis* que, como se vio anteriormente, obtiene mejores resultados que la euclídea. Para cada imagen de ejemplo se obtendrá una lista de imágenes resultado que serán combinadas posteriormente mediante un operador de agregación OWA.

- Algoritmos de expansión de la consulta. Trabaja de forma análoga al algoritmo automático, con la salvedad de que utiliza más imágenes de ejemplo que las proporcionadas en las consultas multimedia. El conjunto de imágenes seleccionado para expandir la consulta se obtiene de las primeras posiciones de la lista de resultados del subsistema TBIR.

- Algoritmo de realimentación por relevancia basado en regresión logística. Este algoritmo es el utilizado finalmente dentro de la propuesta presentada en esta tesis. La descripción detallada del mismo puede encontrarse en el apartado 6.3.

Los resultados obtenidos en esta experimentación no son comparables a los mostrados para el primer conjunto de experimentos, ya que en este caso no se trabaja sobre la colección de imágenes original, sino sobre la versión reducida generada mediante el prefiltro textual.

Se llevan a cabo tres experimentos diferentes sobre cada uno de los tres algoritmos CBIR implementados. La diferencia entre dichos experimentos es el tipo de prefiltrado textual realizado previamente sobre la colección completa de imágenes (cada uno de ellos basado en una configuración TBIR distinta), pero lo interesante en este apartado será comparar la actuación de la recuperación CBIR independientemente del mencionado prefiltrado. Los resultados obtenidos se muestran en la siguiente tabla:

Tabla 7-16. Resultados de algoritmos CBIR (*ImageCLEF* 2010)

	Algoritmo CBIR	**MAP**	**P@10**	**P@20**
Experimento1	Automático	0,1502	**0,3971**	0,3607
	Expansión	0,1525	0,3943	0,3621
	Regresión	**0,1792**	0,3914	**0,3629**
Experimento2	Automático	0,1261	**0,3857**	0,3307
	Expansión	0,1286	0,3829	**0,3386**
	Regresión	**0,1498**	0,3543	0,3250
Experimento3	Automático	0,1089	**0,4043**	0,3357
	Expansión	0,1077	0,3886	0,3307
	Regresión	**0,1285**	0,3614	**0,3379**

Puede observarse cómo el algoritmo CBIR que mejor funciona en términos de MAP es el basado en regresión logística. Para las medidas de evaluación referentes a las precisiones a bajo nivel (*early precision*), P@10 y P@20, los resultados obtenidos son bastante parecidos, aunque parece que el algoritmo automático funciona algo mejor para P@10. Como ya se ha mencionado, los valores de precisión a bajo nivel son de especial interés en sistemas de recuperación de imágenes utilizados por usuarios reales, ya que estos estarán en muchos casos interesados únicamente en las imágenes recuperadas en las primeras posiciones.

Como se verá en el apartado dedicado a la evaluación y análisis de resultados correspondientes al prefiltro textual (apartado 7.3.1), los valores obtenidos para las precisiones bajas (P@5 y P@10) mejoran en muchos casos a los experimentos puramente textuales (TBIR) que dan lugar al prefiltrado utilizado por estos sistemas visuales. Por el

contrario, esto no sucede para la medida de evaluación MAP, para la que los experimentos TBIR son los que mejor rendimiento ofrecen.

Por último, se muestran los resultados obtenidos por el sistema de recuperación CBIR desarrollado para la tarea de recuperación de imágenes de Wikipedia del *ImageCLEF* 2011. El algoritmo utilizado es el de realimentación por relevancia basada en regresión logística (descrito en el apartado 6.3), ya que fue el que mejores resultados obtuvo en la experimentación de la edición 2010 (Tabla 7-16). Los resultados mostrados se refieren a una recuperación visual pura, esto es, sobre la colección de imágenes completa, sin la ayuda del prefiltro textual para acotar la colección.

Tabla 7-17. Resultados CBIR (*ImageCLEF* 2011)

	MAP	**P@10**	**P@20**
CBIR	0.0014	0.0060	0.0040
mejor	0.0044	0.0340	0.0280
media	0.0039	0.0270	0.0245

En la tabla se han añadido tanto los resultados obtenidos por el mejor grupo en la recuperación puramente basada en características visuales, como los valores medios obtenidos por todos los participantes en dicha modalidad. Puede observarse la diferencia entre la aproximación propuesta (CBIR) y los mejores resultados obtenidos por otros participantes, lo que influye también en los resultados tras la fusión multimedia propuesta.

7.2.2 CBIR por tipos de descriptores visuales

En este apartado se definen y ejecutan un conjunto de experimentos que tienen como objetivo el análisis de sistema de recuperación CBIR cuando este trabaja en base a grupos de descriptores visuales independientes, en comparación con cuando lo hace con todos ellos de manera conjunta. El sistema CBIR con el que se evalúan estos aspectos estará basado en el modelo de regresión logística descrito en el apartado 6.3.

Los resultados mostrarán que la mejor opción es contar con todos los descriptores visuales conjuntamente, pero será interesante analizar cuáles de ellos funcionan mejor cuando se aborda la tarea de recuperación.

Se definen cinco grupos de descriptores visuales, entre los que hay dos de color y tres de textura. La nomenclatura para cada uno de ellos es la siguiente:

- ColorG. Histogramas de color global, con un total de 30 descriptores.
- ColorL. Histogramas de color local. La imagen se divide en 4 trozos, y se calculan los histogramas de forma independiente para cada uno de ellos. Se tienen un total de 192 descriptores.
- TexGral0. Descriptor de texturas que calcula las granulometrías usando como elemento estructurante una línea horizontal (0 grados). 31 descriptores.
- TexGral90. Descriptor de texturas basado en granulometrías usando como elemento estructurante una línea vertical (90 grados). 31 descriptores.
- TexSsdl0. Descriptor de texturas que calcula los valores de la *Spatial Size Distribution* usando un elemento estructurante la línea horizontal. 9 descriptores.

La recuperación visual (CBIR) basada en grupos de descriptores independientes no se realizará sobre la colección completa de evaluación utilizada en cada caso, sino que se hará únicamente sobre el conjunto reducido de imágenes seleccionadas mediante el prefiltro textual. La colección de evaluación utilizada es la de imágenes de Wikipedia del *ImageCLEF* 2011.

La siguiente figura muestra gráficamente los experimentos llevados a cabo para evaluar el rendimiento de la recuperación basada en el contenido visual de las imágenes cuando se utilizan diferentes grupos de descriptores. También se incluye el caso en el que se utilizan todos los descriptores visuales disponibles de manera conjunta:

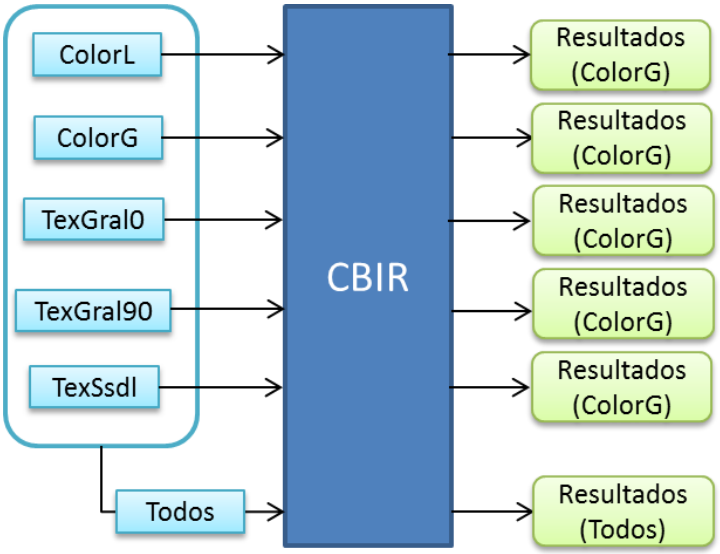

Figura 7.1. Experimentación con grupos de descriptores visuales

Se ejecutarán las 50 consultas disponibles y se generarán, por lo tanto, cinco listas de resultados con las imágenes recuperadas en base a cada conjunto de descriptores. Los resultados obtenidos tras la evaluación de estas listas puede verse en la Tabla 7-18, donde se muestran los valores de MAP y precisiones a bajo nivel obtenidos gracias a cada uno de los grupos de descriptores visuales definidos. Se incluyen también los valores obtenidos cuando se trabaja conjuntamente con todos los descriptores para facilitar la comparación de los resultados:

Tabla 7-18. Resultados CBIR por grupos de descriptores visuales

Descriptores Visuales	MAP	P@5	P@10	P@20
ColorG	0,0348	0,0520	0,0540	0,0480
ColorL	0,0552	**0,1000**	0,0860	0,0710
TexGral0	0,0217	0,0320	0,0300	0,0320
TexGral90	0,0221	0,0200	0,0340	0,0300
TexSsdl0	0,0235	0,0160	0,0220	0,0260
Todos	**0,0618**	0,0880	**0,0880**	**0,0910**

Observando los valores obtenidos para MAP, el grupo de descriptores que mejores resultados obtiene es el correspondiente a ColorL (color local), exceptuando cunado se trabaja sobre el conjunto global (Todos). Esto puede ser debido a las características propias de los

descriptores utilizados en ese grupo, o más probablemente al número de ellos (192). Otro dato a resaltar es que la P@5 obtenida en base al grupo de descriptores ColorL es mayor incluso que la obtenida cuando se trabaja con el conjunto total de descriptores. En cualquier caso, los resultados obtenidos cuando se trabaja con todos los descriptores disponibles de manera conjunta son superiores a los obtenidos en los experimentos independientes (salvo en P@5 para la que ColorL funciona ligeramente mejor).

Esta serie de experimentos por grupos de descriptores visuales independientes serán combinados, cada uno de ellos, con los resultados textuales siguiendo la estrategia LSMF propuesta en esta tesis. Dicha experimentación junto con los resultados obtenidos y su análisis se incluirá en el apartado 7.3.6

7.3 Fusión Multimedia

Los experimentos en este apartado están destinados a evaluar el comportamiento de la propuesta de Fusión Multimedia Semántica Tardía (LSMF, *Late Semantic Multimedia Fusion*), cuando es aplicada en una tarea de recuperación multimedia de imágenes. La primera parte de esta experimentación estará dedicada a la evaluación del prefiltro textual, a continuación se pasará a evaluar y analizar el comportamiento de diferentes algoritmos de fusión (descritos en la sección 6.4), que completarán la técnica de LSMF. Más adelante se incluye una experimentación relacionada con la normalización de los *scores* (valores de relevancia) de las listas de resultados de los subsistemas implicados (textual y visual), para finalizar con un análisis del funcionamiento de la propuesta si CBIR trabajara con grupos de descriptores visuales (como se explicó en la sección 7.2.2).

7.3.1 Prefiltro Textual

Usar los resultados obtenidos por el sistema de recuperación textual (TBIR) para reducir la colección de imágenes original y que CBIR trabaje únicamente sobre aquellas imágenes que, según las características textuales, tengan alguna relación semántica con la necesidad de información expresada por los usuarios en las consultas multimedia permite, además proporcionar contraejemplos (imágenes de ejemplos negativos) para los algoritmos visuales.

Se muestra a continuación la experimentación llevada a cabo con el objetivo de comparar el rendimiento del sistema de recuperación visual cuando se introduce la fase de prefiltrado y

cuando no (recuperación visual pura). Los experimentos son desarrollados sobre la colección de imágenes de Wikipedia, utilizando las consultas de la edición de 2011 de *ImageCLEF*. Se incluye también un análisis de la calidad de la versión reducida generada a partir del conjunto de imágenes original.

Tabla 7-19. Mejora CBIR con Prefiltro

	MAP	**P@5**	**P@10**	**P@20**
CBIR	0,0014	0,0060	0,0060	0,0040
CBIR+prefiltro	**0,0618**	0,0880	0,0880	0,0910

La tabla anterior muestra cómo los resultados de los experimentos visuales son mejorados notablemente gracias a la fase de prefiltrado textual, sin olvidar cómo aumenta la eficiencia por la reducción del conjunto de imágenes previa al proceso CBIR. Se pasa de una precisión media (MAP) del 0.14% cuando el sistema CBIR trabaja directamente sobre la colección original, al 6.18% cuando el sistema TBIR prefiltra o acota dicha colección. Igualmente, los valores de precisión a bajo nivel (*early precision*) aumentan considerablemente cuando se incluye la etapa de prefiltrado. La siguiente figura muestra visualmente esta importante mejora:

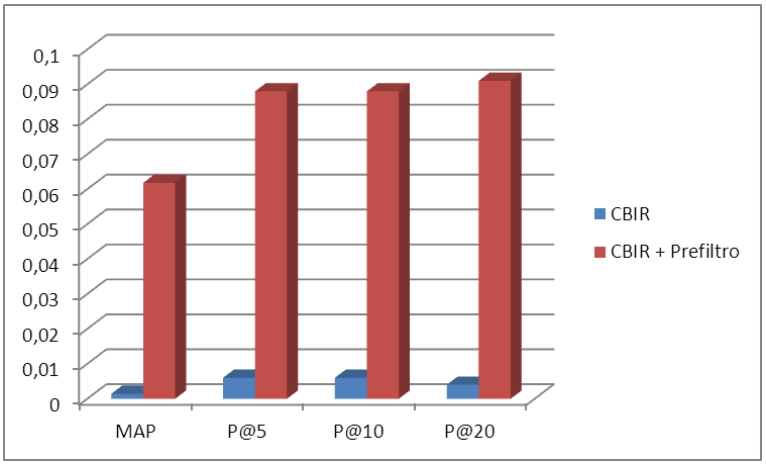

Figura 7.2. Mejora CBIR con Prefiltro (*topics2011*)

A continuación se analiza la actuación del prefiltro sobre la colección de imágenes de Wikipedia de *ImageCLEF*, utilizando las consultas propuestas por la organización tanto para la edición del 2010 (70 consultas) como para la del 2011 (50 consultas). También se incluye

el mismo análisis sobre el conjunto de consultas formado por las proporcionadas para ambas ediciones (2010+2011). La tabla 7-18 indica el número de imágenes presentes en la colección original y en cada una de las subcolecciones generadas por el prefiltro textual, así como el porcentaje de reducción logrado, y la cobertura (proporción de imágenes relevantes que superan el prefiltro en relación al número de imágenes relevantes total en la colección) mantenida con las versiones reducidas (y relacionadas semánticamente con las consultas según la información textual) de la colección:

Tabla 7-20. Reducción de la colección y cobertura tras Prefiltro

Colección	Tamaño (nº imágenes)	Reducción	Imágenes relevantes	Cobertura
Original	237.434	-	17.660 (2010) 3.440 (2011)	100 %
Tras Prefiltro (*2010*)	4.507	98,10 %	13.914	78,79 %
Tras Prefiltro (*2011*)	5.490	97,69 %	2.854	82,97 %
Tras Prefiltro (*2010+2011*)	4.916	97,93 %	16.768	79,47 %

Como puede observarse, la reducción de la colección de imágenes, casi un 98% de media por consulta, no afecta de la misma manera a la cobertura mantenida por el subconjunto de imágenes que superan el prefiltro textual ($\approx$ 80%). Esto es importante, ya que se introducen ventajas como la mejora de los resultados visuales, la escalabilidad de la tarea sobre grandes colecciones, y la posibilidad de utilizar contraejemplos.

En el Capítulo 5 (Figura 5.6) ya se mostraron gráficamente estos beneficios, así como un ejemplo práctico en base una de las consultas multimedia propuestas en la edición de 2011 del *ImageCLEF* (Figura 5.7), en el que se muestra la mejora de los resultados CBIR cuando se introduce la etapa de prefiltrado textual (pasando de 0 a 7 imágenes relevantes recuperadas entre los primeros 15 resultados).

7.3.2 Fusión Tardía

Este apartado está dedicado a la experimentación relacionada con la fusión multimedia a nivel de decisiones o tardía (*late fusion*) llevada a cabo entre las listas de resultados (conjunto de imágenes ordenadas por relevancia) obtenidas desde el sistema textual (TBIR) y desde el

visual puro (CBIR sin prefiltro). Los algoritmos implementados de fusión tardía que se evaluarán han sido descritos detalladamente en el apartado 6.4.

Los resultados obtenidos en base a estos algoritmos de fusión tardía, utilizando la colección de imágenes de Wikipedia y las consultas proporcionadas en la edición de 2011 de *ImageCLEF*, se muestran en las siguientes tablas. Se incluye en ellas, para facilitar el análisis, los valores de evaluación obtenidos por los experimentos TBIR y CBIR (*baselines*). Los experimentos monomodales fusionados son los siguientes:

- TBIR: experimento textual con la configuración descrita en el apartado 7.1.5. No se aplica ninguna técnica de normalización de *scores*.

- CBIR: experimento visual en base al esquema descrito en el apartado 6.3. Este sistema trabaja sobre la colección completa original (sin prefiltrar y sin contraejemplos), y no aplica ningún proceso de normalización a sus valores de relevancia o *scores*.

En primer lugar se plantean diversos experimentos enfocados a analizar el comportamiento de cada uno de los algoritmos de fusión tardía implementados. Aquellos que incorporan algún parámetro de configuración son evaluados aplicando distintos valores a dicho parámetros. Los algoritmos de este tipo son *OWA*, *Enrich* y *FilterN*.

Para el algoritmo de fusión OWA se utilizan distintos valores para la medida *orness* que, como se describe en el apartado 6.4, indica el grado en que la agregación se asemeja a una operación de tipo OR. Los resultados de evaluación obtenidos para valores de *orness* desde 0 hasta 1 (con incrementos de 0.1) se muestran en la siguiente tabla. Los resultados resaltados en negrita corresponden a aquellos experimentos en los que se ha mejorado el rendimiento de la aproximación basada en texto (TBIR), que es la aproximación monomodal que mejor rendimiento ofrece.

Tabla 7-21. Resultados Fusión Tardía - *OWA*

		MAP	Mejora vs TBIR	P@5	Mejora vs TBIR	P@10	Mejora vs TBIR	P@20	Mejora vs TBIR
TBIR		0,3044	-	0,5600	-	0,5060	-	0,4040	-
CBIR		0,0014	-	0,0080	-	0,0060	-	0,0040	-
OWA	**min**	0,1837	-39,65%	0,3000	-46,43%	0,2700	-46,64%	0,2540	-37,13%
	orness01	**0,3068**	+0,79%	**0,5760**	+2,86%	0,5060	0,00%	**0,4140**	+2,48%
	orness02	**0,3254**	+6,90%	**0,6200**	+10,71%	**0,5440**	+7,51%	**0,4250**	+5,20%
	orness03	**0,3274**	+7,56%	**0,6200**	+10,71%	**0,5320**	+5,14%	**0,4310**	+6,68%
	orness04	**0,3240**	+6,44%	**0,5880**	+5,00%	**0,5260**	+3,95%	**0,4220**	+4,46%
	avg	**0,3158**	+3,75%	**0,5840**	+4,29%	**0,5220**	+3,16%	**0,4220**	+4,46%
	orness06	**0,3093**	+1,61%	**0,5840**	+4,29%	**0,5140**	+1,58%	**0,4130**	+2,23%
	orness07	0,3001	-1,41%	**0,5760**	+2,86%	0,5060	0,00%	**0,4080**	+0,99%
	orness08	0,2885	-5,22%	**0,5680**	+1,43%	0,5000	-1,19%	0,4000	-0,99%
	orness09	0,2787	-8,44%	**0,5720**	+2,14%	0,5020	-0,79%	0,3960	-1,98%
	max	0,2688	-11,70%	0,5600	0,00%	0,5000	-1,19%	0,3980	-1,49%

Los resultados obtenidos se muestran gráficamente en la siguiente figura, con el objetivo de facilitar su análisis y explicación.

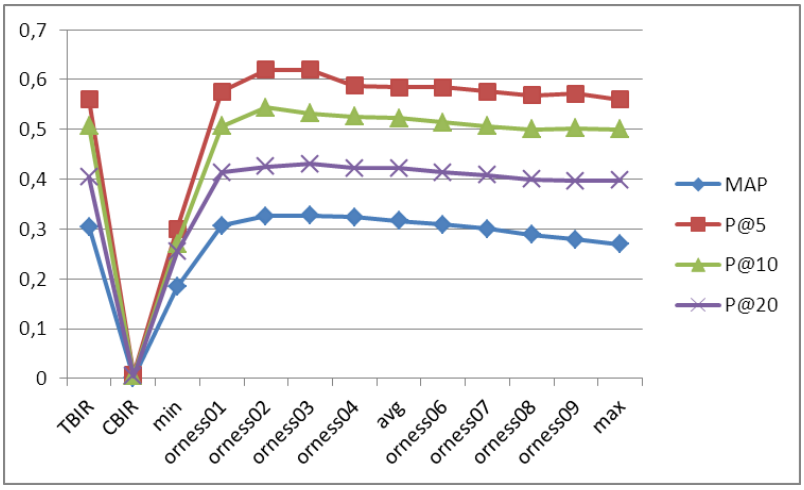

Figura 7.3. Gráfica comparativa Fusión Tardía – *OWA*

Analizando los valores obtenidos, puede observarse cómo muchos de los resultados obtenidos mejoran el rendimiento de la aproximación basada únicamente en texto (TBIR). Los resultados obtenidos a partir de las características visuales (CBIR) son siempre mejorados. Por tanto, se confirma el hecho de que la colaboración entre ambas modalidades (textual y visual) es capaz explotar la complementariedad existente entre ellas para mejorar los resultados monomodales.

Las mejoras se observan para valores bajos (exceptuando el 0) de la medida *orness* (desde 0.1 hasta 0.6), dándose las mejoras más significativas para los valores 0.2 y 0.3, para los que se consigue mejorar el MAP del *baseline* textual entre un 5% y un 10%, y P@5 en más del 10%. Para los casos en los que se produce una mejora en relación a los resultados textuales, esta se da para todas las medidas de evaluación analizadas: tanto para el promedio de las precisiones medias (MAP), como para las precisiones bajas o *early precisions* (P@5, P@10 y P@20).

Se analizan ahora los resultados obtenidos para distintas configuraciones del algoritmo de fusión tardía *FilterN*. La evaluación llevada a cabo para distintos valores del umbral de filtrado N se muestra en la siguiente tabla, junto con los valores de los *baselines* de texto (TBIR) e imagen (CBIR). Los valores resaltados en negrita se corresponden con los mejores resultados obtenidos para cada una de las medidas de evaluación mostradas (MAP, P@5, P@10 y P@20).

Tabla 7-22. Resultados Fusión Tardía - *FilterN*

		MAP	**P@5**	**P@10**	**P@20**
TBIR		0,3044	0,5600	0,5060	0,4040
CBIR		0,0014	0,0080	0,0060	0,0040
FilterN	**N=1000**	0,0311	0,1600	0,1080	0,0700
	N=1500	0,0392	0,2200	0,1460	0,0900
	N=2000	0,0396	0,2440	0,1660	0,1050
	N=2500	0,0429	0,2520	0,1780	0,1130
	N=5000	0,0673	0,3360	0,2480	0,1700
	N=10000	**0,0879**	**0,3840**	**0,3060**	**0,2320**

Los resultados obtenidos tras la fusión multimedia no mejoran en ningún caso los obtenidos por el experimento basado únicamente en el texto de las imágenes, ya que los bajos resultados del sistema CBIR afectan de manera negativa a la fusión. La intención de este algoritmo, como se detalla en el apartado 6.4, es eliminar de la lista de resultados textuales aquellos que no se son recuperados entre las primeras N posiciones por el sistema CBIR. Al trabajar con la colección de imágenes original, los resultados visuales son de una precisión baja y, por lo tanto, no ayudan a eliminar imágenes no relevantes (falsos positivos) de la lista de resultados obtenida desde el sistema TBIR. Al contrario, eliminan un gran número de imágenes

relevantes (verdaderos positivos) y, de este, modo afectan negativamente al rendimiento de la fusión multimedia.

La tabla de resultados para el algoritmo *Enrich* es la que se muestra a continuación. En ella se evalúa el rendimiento del algoritmo para diferentes ponderaciones del valor de relevancia de la lista secundaria, esto es, la lista CBIR, tal y como se describe en el apartado 6.4. Los valores resaltados en negrita corresponden a aquellos casos en los que la configuración del algoritmo *Enrich* mejora los resultados monomodales de texto.

Tabla 7-23. Resultados Fusión Tardía - *Enrich*

		MAP	**P@5**	**P@10**	**P@20**
TBIR		0.3044	0.5600	0.5060	0.4040
CBIR		0.0014	0.0080	0.0060	0.0040
Enrich	**x1**	**0.3060**	0.5600	**0.5080**	0.4010
	x1.25	**0.3061**	0.5600	**0.5080**	0.4030
	x1.5	**0.3061**	0.5600	**0.5080**	0.4030
	x2	**0.3061**	0.5600	**0.5080**	0.4030
	x5	**0.3053**	0.5560	**0.5100**	0.4010

Puede observarse cómo los resultados obtenidos son bastante similares a los correspondientes al *baseline* textual. La influencia del valor de relevancia o *score* de la lista de imágenes recuperadas por el sistema CBIR (lista secundaria) parece ser escasa o nula en el *score* final tras el proceso de fusión. Esto puede ser debido a la mayor magnitud del rango de valores obtenidos para la lista de resultados textual en comparación con la visual, ya que ningún proceso de normalización es aplicado a las listas originales. Este aspecto se analizará más adelante, dentro de los experimentos de fusión multimedia semántica tardía (LSMF), donde se estudiará el comportamiento de la fusión según distintas técnicas de normalización.

Los resultados correspondientes a los algoritmos de fusión *Product* y *MaxMerge* (algoritmos sin parámetros de configuración) son mostrados conjuntamente (Tabla 7-24) con las mejores configuraciones de los algoritmos ya vistos, con el objetivo de comparar el rendimiento de todos en la tarea de fusión multimedia tardía o a nivel de decisiones.

Tabla 7-24. Resultados algoritmos Fusión Tardía

	MAP	Mejora vs TBIR	P@5	Mejora vs TBIR	P@10	Mejora vs TBIR	P@20	Mejora vs TBIR
TBIR	0,3044	-	0,5600	-	0,5060	-	0,4040	-
CBIR	0,0014	-	0,0080	-	0,0060	-	0,0040	-
Product	**0,3271**	+7,46%	**0,6160**	+10,00%	**0,5480**	+8,30%	**0,4380**	+8,42%
MaxMerge	0,2688	-11,70%	0,5600	0,00%	0,5000	-1,19%	0,3980	-1,49%
OWA	**0,3254**	+6,90%	**0,6200**	+10,71%	**0,5440**	+7,51%	**0,4250**	+5,20%
FilterN	0,0879	-71,12%	0,3840	-31,43%	0,3060	-39,53%	0,2320	-42,57%
Enrich	**0,3061**	+0,56%	0,5600	0,00%	**0,5080**	+0,40%	0,4030	-0,25%

Estos resultados se muestran gráficamente en la siguiente figura para facilitar su análisis:

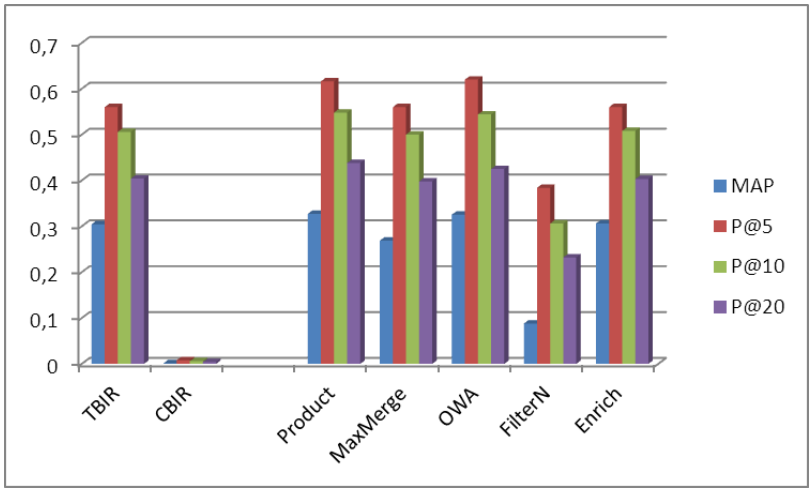

Figura 7.4. Gráfica comparativa algoritmos Fusión Tardía

Se observa que los algoritmos que ofrecen un mejor rendimiento son el *Product* y el *OWA* (con $orness = 0{,}2$). Los resultados obtenidos por estos algoritmos superan al resto tanto en valores de MAP como en precisiones bajas (P@5, P@10 y P@20). De entre estos dos algoritmos, parece que el rendimiento del *Product* es ligeramente superior al del OWA, con mejoras cercanas al 10% en todas las medidas de evaluación analizadas.

En la siguiente sección se analizará el comportamiento de estos mismos algoritmos dentro de la estrategia de fusión multimedia semántica tardía (LSMF) propuesta en esta tesis. Se analizará si los resultados siguen mejorando los *baselines* monomodales cuando se reduce significativamente la colección de trabajo y, por lo tanto, el tiempo de computación y proceso, lo que haría escalable la tarea de recuperación sobre grandes colecciones multimedia.

7.3.3 Fusión Multimedia Semántica Tardía (LSMF)

Este apartado está dedicado a evaluar y analizar el rendimiento de la estrategia de fusión multimedia propuesta en esta tesis (LSMF). Esta estrategia se basa en la inclusión de un prefiltro (basado en la información textual) que acota la colección, eliminando de ella aquellas imágenes que no guarden relación semántica alguna con las consultas (aprovechando el mayor nivel semántico de la modalidad textual en relación a la visual, como se muestra en la Figura 2.5). Tras la aplicación de este prefiltro, la recuperación visual (CBIR) trabajará sobre la colección prefiltrada, lo que aumenta su eficiencia (Figura 5.6) y hace escalable la tarea de recuperación multimedia sobre colecciones grandes. Una vez que ambos sistemas monomodales (TBIR y CBIR) obtienen su lista de imágenes recuperadas, estas son combinadas haciendo uso de los algoritmos de fusión tardía descritos en el apartado 6.4.

La colección de evaluación es la utilizada en las ediciones de 2010 y 2011 de la tarea de recuperación de imágenes de Wikipedia del foro *ImageCLEF*. Las listas de resultados que se van a fusionar son las siguientes:

- TBIR: experimento textual con la configuración descrita en el apartado 7.1.5. No se aplica ninguna técnica de normalización de *scores*.

- CBIR: experimento visual en base al esquema descrito en el apartado 6.3. Este sistema trabaja sobre la versión simplificada de la colección (tras el prefiltrado textual), y no aplica ningún proceso de normalización a sus valores de relevancia o *scores*.

Se empieza analizando el comportamiento del algoritmo de fusión OWA, dentro de la estrategia LSMF, con diferentes configuraciones para el valor de la medida *orness* aplicada (descrita en el apartado 6.4). Se lanzan experimentos con valores de *orness* desde 0 (operación tipo AND) hasta 1 (OR) con incrementos de 0.1. Los resultados obtenidos se muestran en la siguiente tabla, donde también se incluyen los correspondientes a los experimentos monomodales fusionados (TBIR y CBIR). Los resultados resaltados en negrita corresponden a aquellos experimentos en los que se ha mejorado el rendimiento de la aproximación basada en texto (TBIR), que es la aproximación monomodal que mejor rendimiento ofrece.

Tabla 7-25. Resultados LSMF - *OWA*

		MAP	Mejora vs TBIR	P@5	Mejora vs TBIR	P@10	Mejora vs TBIR	P@20	Mejora vs TBIR
TBIR		0,3044	-	0,5600	-	0,5060	-	0,4040	-
CBIR		0,0618		0,088		0,088		0,091	
OWA	min	0,1777	-41,62%	0,3520	-37,14%	0,3160	-37,55%	0,2660	-34,16%
	orness01	**0,3103**	1,94%	**0,6360**	13,57%	**0,5280**	4,35%	**0,4290**	6,19%
	orness02	**0,3369**	10,68%	**0,6600**	17,86%	**0,5660**	11,86%	**0,4450**	10,15%
	orness03	**0,3371**	10,74%	**0,6440**	15,00%	**0,5540**	9,49%	**0,4480**	10,89%
	orness04	**0,3319**	9,03%	**0,6200**	10,71%	**0,5420**	7,11%	**0,4330**	7,18%
	avg	**0,3233**	6,21%	**0,6000**	7,14%	**0,5240**	3,56%	**0,4350**	7,67%
	orness06	**0,3174**	4,27%	**0,5840**	4,29%	**0,5200**	2,77%	**0,4240**	4,95%
	orness07	**0,3108**	2,10%	**0,5800**	3,57%	**0,5080**	0,40%	**0,4130**	2,23%
	orness08	**0,3049**	0,16%	**0,5720**	2,14%	0,5060	0,00%	**0,4080**	0,99%
	orness09	0,2994	-1,64%	**0,5720**	2,14%	0,5000	-1,19%	0,4020	-0,50%
	max	0,2933	-3,65%	0,5600	0,00%	0,5000	-1,19%	0,3980	-1,49%

Los resultados obtenidos se muestran gráficamente en la siguiente figura, con el objetivo de facilitar su análisis y explicación.

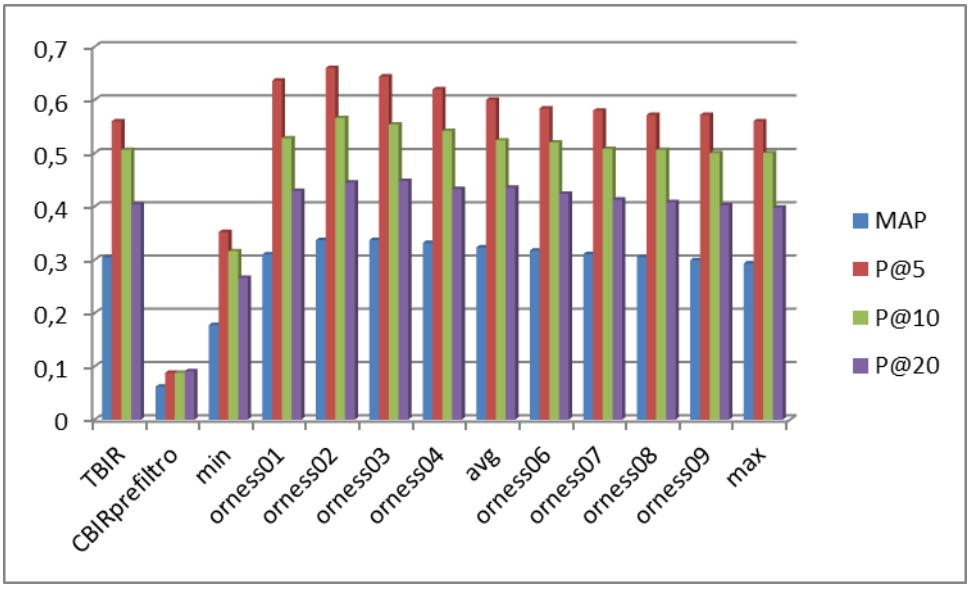

Figura 7.5. Comparativa Fusión Semántica Multimedia Tardía – *OWA*

Se observa que las configuraciones del algoritmo OWA que mejor funcionan dentro de la técnica de LSMF son las correspondientes a *orness02* y *orness03*. Se trata de los mismos casos para los que la aproximación de fusión tardía clásica (sin prefiltro textual semántico) obtenía sus mejores resultados. Ahora dentro de la LSMF los resultados fusionados con el

algoritmo OWA mejoran los *baselines* monomodales para casi todos los valores de *orness*. Solo las configuraciones correspondientes a los extremos no mejoran alguna de las medidas de evaluación analizadas (MAP, P@5, P@10, P@20). Estos valores extremos (*orness00* o min y *orness10* o max) se corresponden con los operadores lógicos AND y OR respectivamente.

En resumen, el rendimiento del algoritmo OWA dentro de la estrategia de combinación LSMF es capaz de mejorar los resultados tanto del *baseline* textual (TBIR) como del visual (CBIR), y esto simplificando significativamente la colección de búsqueda para el proceso visual, con las ventajas de escalabilidad, tiempo de cómputo, disponibilidad de contraejemplos, etc.

Para el algoritmo de fusión *FilterN* se experimenta con distintos valores para el parámetro N, correspondiente al número de elementos recuperados (imágenes) por el sistema visual en sus primeras posiciones que son tenidos en cuenta para no filtrar/eliminar de la lista de resultados textual, tal y como se detalla en el apartado 6.4. Los resultados obtenidos tras la evaluación del algoritmo con valores de $N = 1000, 1500, 2000, 2500, 5000, 10000$, son los mostrados en la siguiente tabla, donde se resaltan en negrita los valores que superan los resultados obtenidos por el sistema TBIR para cada una de las medidas analizadas (MAP, P@5, P@10 y P@20):

Tabla 7-26. Resultados LSMF - *FilterN*

		MAP	**P@5**	**P@10**	**P@20**
TBIR		0,3044	0,5600	0,5060	0,4040
CBIR		0,0014	0,0080	0,0060	0,0040
FilterN	N=500	0,1821	0,5160	0,4460	0,3450
	N=750	0,2127	0,5280	0,4780	0,3760
	N=1000	0,2395	**0,5720**	0,5000	0,4000
	N=1500	0,2900	**0,5680**	0,5040	0,4030
	N=2000	**0,3065**	**0,5640**	**0,5100**	0,4040
	N=2500	**0,3066**	**0,5640**	**0,5100**	0,4040
	N=5000	**0,3056**	**0,5640**	**0,5080**	0,4040
	N=10000	**0,3048**	0,5600	0,5060	0,4040

Analizando los resultados se observa que estos son bastante parecidos en general a los obtenidos por el *baseline* textual (TBIR). Solo en los casos correspondientes a valores de N muy restrictivos (1000 y 1500) el MAP obtenido es más bajo que el de TBIR, que son precisamente los casos para los que se obtienen mejores resultados de P@5.

La razón para unos valores de evaluación tan semejantes está en que la colección sobre la que trabaja el sistema CBIR es la versión simplificada (tras prefiltrar textualmente). Por lo tanto, los resultados visuales serán una reordenación o *reranking* de la lista textual. Entonces, si de la lista de resultados textuales se eliminan aquellas imágenes que no hayan sido recuperadas entre las N primeras posiciones de la lista visual (*reranking*), se obtiene una lista fusionada resultante bastante similar a la original (lista textual).

Se puede observar cómo según aumenta el valor de N, los resultados obtenidos tienden a igualarse con los correspondientes a los obtenidos por la lista textual (TBIR), en base a lo explicado en el párrafo anterior. En el caso de utilizar un valor de N superior al máximo número de resultados recuperados por alguna consulta ($N > \max_{0 \le i \le |Q|} Q_i$, siendo |Q| el número total de consultas, y Q_i el número de resultados recuperados para la consulta i) , el algoritmo *FilterN* no tendría efecto y, por lo tanto, la lista fusionada será la misma que la original (ya que no se filtrará/eliminará ninguna imagen). Para el caso de la colección y el conjunto de consultas sobre el que aquí se experimenta, este valor se correspondería con la consulta número 119 que recupera 32.959 resultados). Solo para algunos de los casos más restrictivos (valores pequeños de N) se consiguen mejoras en los valores de precisión más baja (P@5). En cualquier caso, el valor de N tampoco puede disminuirse demasiado porque incluso este valor empezaría a empeorar. La figura siguiente muestra un gráfico sobre la evolución de los valores de evaluación en función de N:

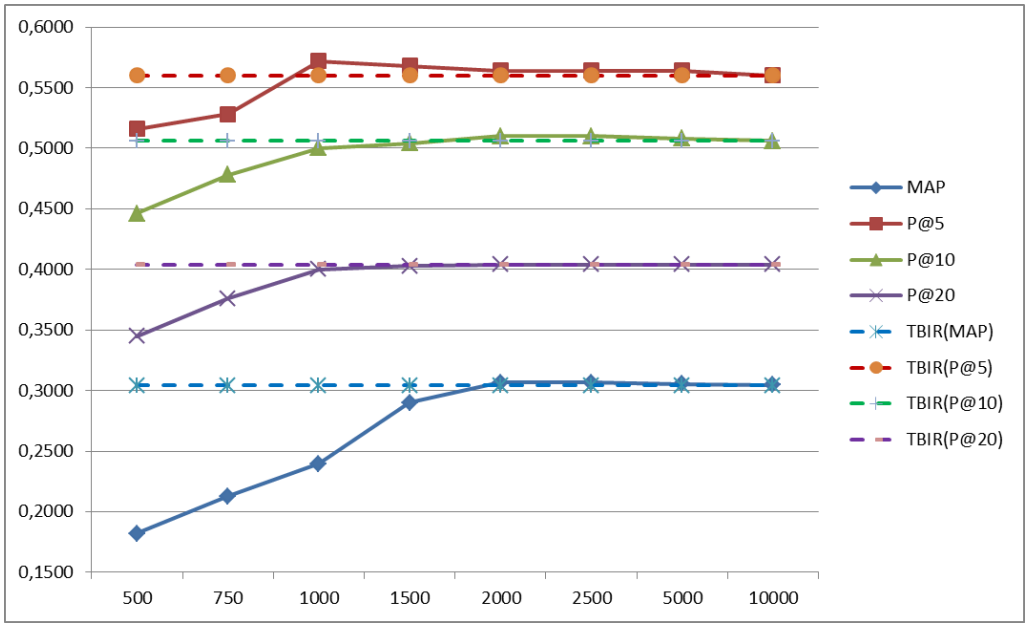

Figura 7.6. Resultados LSMF - FilterN

Puede verse en la gráfica lo que se comentaba en el párrafo anterior: los resultados obtenidos se van igualando a los correspondientes al experimento textual (TBIR) según aumenta el valor del parámetro N. Para algunos de los valores pequeños de N (por ejemplo N=1000) se produce una ligera mejora en los resultados correspondientes a la medida de evaluación P@5, la cual va desapareciendo según aumenta el valor de N. Valores más bajos de N resultan demasiado restrictivo y la mejora apreciada en P@5 desaparece igualmente.

En resumen, el algoritmo de fusión *FilterN* utilizado dentro de la estrategia de combinación multimedia LSMF no obtiene mejoras suficientemente relevantes en relación a las obtenidas por el experimento monomodal textual. Se observa que a partir de valores de N por encima de 2000 los resultados obtenidos son ligeramente superiores. En cualquier caso, las mejoras no son suficientemente significativas como para justificar la aplicación del algoritmo. Comparando el rendimiento obtenido por *FilterN* como parte de LSMF y con el escenario de fusión tardía básica, se comprueba que el funcionamiento es mejor, ya que se trabaja sobre la versión simplificada de la colección, evitando los inconvenientes descritos en el apartado 7.3.2 relacionados con la baja precisión del sistema CBIR cuando trabaja sobre la colección original.

La Tabla 7-27 muestra los resultados obtenidos para los experimentos realizados con el algoritmo de fusión *Enrich*, dentro de la estrategia LSMF. En ella se incluyen resultados de diferentes configuraciones para el algoritmo, probando con distintos valores de enriquecimiento tal y como se describe en el apartado 6.4: el incremento aplicado al *score* de la lista principal (TBIR) es multiplicado por los valores mostrados en la tabla (1, 1.25, 1.5, 2 y 5). Se resaltan en negrita los valores que mejoran el *baseline* textual.

Tabla 7-27. Resultados LSMF - *Enrich*

		MAP	P@5	P@10	P@20
TBIR		0,3044	0,5600	0,5060	0,4040
CBIR		0,0014	0,0080	0,0060	0,0040
Enrich	**x1**	**0,3079**	**0,5640**	**0,5080**	0,4050
	x1.25	**0,3074**	0,5560	**0,5080**	0,4060
	x1.5	**0,3079**	0,5600	0,5040	**0,4070**
	x2	0,3043	0,5520	0,5000	0,4050
	x5	0,2766	0,4720	0,4500	0,3900

Puede observarse cómo los resultados son bastante parecidos a los obtenidos por el sistema TBIR. Esto es lo mismo que sucedía dentro del esquema clásico de fusión tardía (apartado 7.3.2. La posible causa puede ser la diferencia entre los rangos de valores de los *scores* de los resultados textuales y visuales. Esto se comprobará más adelante, cuando se analicen los experimentos referentes a la estrategia LSMF utilizando listas de resultados con valores de relevancia o *scores* normalizados (apartado 7.3.4), esto es, dentro del mismo rango de valores.

Los resultados correspondientes a los algoritmos de fusión *Product* y *MaxMerge* (algoritmos sin parámetros de configuración) son mostrados conjuntamente con las mejores configuraciones del resto de algoritmos anteriores. La columna “Mejora VS TBIR” indica el porcentaje de mejora relativa obtenido por el algoritmo analizado en relación al *baseline* textual.

Tabla 7-28. Resultados algoritmos LSMF

	MAP	Mejora vs TBIR	P@5	Mejora vs TBIR	P@10	Mejora vs TBIR	P@20	Mejora vs TBIR
TBIR	0,3044	-	0,5600	-	0,5060	-	0,4040	-
CBIR+STP	0,0618	-	0,0880	-	0,0880	-	0,0910	-
Product	**0,3400***	11,70%	**0,6600**	17,86%	**0,5540**	9,49%	**0,4550**	12,62%
MaxMerge	0,2933	-3,65%	0,5600	0,00%	0,5000	-1,19%	0,3980	-1,49%
OWA	**0,3369**	10,68%	**0,6600**	17,86%	**0,5660**	11,86%	**0,4450**	10,15%
FilterN	0,2395	-21,32%	**0,5720**	2,14%	0,5000	-1,19%	0,4000	-0,99%
Enrich	**0,3079**	1,15%	**0,5640**	0,71%	**0,5080**	0,40%	**0,4050**	0,25%

(*) Mejora estadísticamente significativa en relación al experimento TBIR

Estos resultados se muestran gráficamente en la siguiente figura para facilitar su análisis:

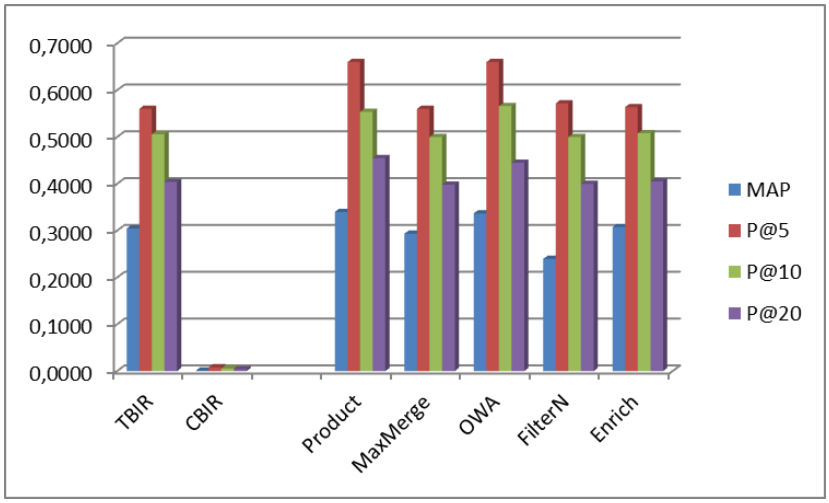

Figura 7.7. Comparativa algoritmos LSMF

Puede observarse cómo el rendimiento de los algoritmos *Product* y *OWA* resaltan por encima del resto, superándose en ambos casos los resultados del *baseline* textual en más de 10% para casi todas las medidas de evaluación analizadas. El valor de MAP más alto obtenido corresponde al algoritmo *Product*, que alcanza casi un 12% de mejora relativa (11,70%). Para el caso de las medidas referentes a las precisiones a bajo nivel (*early precisions*) cabe destacar el incremento de casi un 18% con respecto a TBIR de los algoritmos *Product* y *OWA*. Recordar la importancia de este tipo de precisiones que, como se dijo en los apartados relacionados con el estado del arte, miden la calidad de las imágenes recuperadas en las primeras posiciones de la lista de resultados que, en la mayoría de los casos, serán las únicas que mirarán los usuarios reales en un sistema de recuperación multimedia. También resaltar el

hecho de que todas estas mejoras son conseguidas dentro del esquema de fusión propuesto en esta tesis (LSMF), que reduce significativamente el tiempo de proceso del sistema visual y permite la escalabilidad de la tarea de recuperación sobre colecciones multimedia de gran tamaño.

La comparación que se muestra a continuación entre el rendimiento de la fusión con LSMF y con un esquema cásico de fusión tardía (LF), muestra la conveniencia de aplicar la alternativa propuesta en esta tesis.

Tabla 7-29. Comparación de algoritmos con LSMF y Fusión Tardía clásica (LF)

		MAP	P@5	P@10	P@20
Product	LF	0,3271	0,6160	0,5480	0,4380
	LSMF	**0,3400**	**0,6600**	**0,5540**	**0,4550**
MaxMerge	LF	0,2688	0,5600	0,5000	0,3980
	LSMF	**0,2933**	0,5600	0,5000	0,3980
OWA	LF	0,3254	0,6200	0,5440	0,4250
	LSMF	**0,3369**	**0,6600**	**0,5660**	**0,4450**
FilterN	LF	0,0879	0,3840	0,3060	0,2320
	LSMF	**0,2395**	**0,5720**	**0,5000**	**0,4000**
Enrich	LF	0,3061	0,5600	0,5080	0,4030
	LSMF	**0,3079**	**0,5640**	0,5080	**0,4050**

Puede observarse cómo en todos los casos la aproximación basada en LSMF mejora o iguala los resultados obtenidos por la fusión tardía clásica que trabaja sobre la colección completa. A parte de la considerable simplificación de los procesos de recuperación visual y fusión multimedia, los valores de precisión (tanto general con MAP, como a bajo nivel con P@5, P@10 y P@20) son mejorados gracias a la calidad del espacio de búsqueda reducido generado por el prefiltro textual semántico.

7.3.4 Comparación con otras aproximaciones de fusión

La Tabla 7-30 resume los resultados obtenidos por los experimentos correspondientes a la estrategia de fusión propuesta (LSMF) en comparación con las mejores aproximaciones y la media de todos los grupos participantes en *ImageCLEF* tanto para la edición de 2010 como para la de 2011 (que son las dos últimas ediciones en las que se utilizó la colección de evaluación empleada). Se describen antes los experimentos comparados:

- **LSMF** (*Late Semantic Multimedia Fusion*): el algoritmo de fusión tardía empleado es el *Product*, con prefiltrado textual para CBIR.

- **media**: resultados correspondientes al valor medio obtenido por todos los experimentos presentados a cada una de las ediciones de la competición. Se tienen en cuenta únicamente aquellas aproximaciones que utilizan tanto la información textual como la visual, esto es, experimentos multimodales. Para la edición de 2010 se presentaron un total de 9 grupos de investigación distintos con 72 experimentos, mientras que en la de 2011 fueron 8 grupos con 57 experimentos presentados.

- **mejor**: experimentos con los mejores resultados de entre todos los presentados a cada una de las ediciones de la tarea.

Tabla 7-30. Comparación LSMF en *ImageCLEF* 2010 y 2011

		MAP	**P@10**	**P@20**
ImageCLEFwiki2010	**LSMF**	0,3111	0,5929	0,5479
	media	0,1387	0,3701	0,3293
	mejor	0,2765	0,5814	0,5193
ImageCLEFwiki2011	**LSMF**	0,3400	0,5540	0,4550
	media	0,2558	0,4542	0,3678
	mejor	0,3880	0,6320	0,5100
ImageCLEFwiki 2010+2011	**LSMF**	0,3231	0,5767	0,5092
	media	0,1875	0,4051	0,3453
	mejor	0,3230	0,6025	0,5154

Puede observarse cómo la aproximación LSMF propuesta se encuentra al nivel de la mejor de todas las presentadas en el *ImageCLEF* durante sus dos últimos años. La medida P@5 no se incluye en la tabla porque no es proporcionada por la organización para el resto de grupos. Haciendo la evaluación global sobre el conjunto total de consultas propuestos para las ediciones de 2010 y 2011 (que utilizan la misma colección de evaluación, descrita en el apartado 4.2.1.3), los resultados obtenidos por nuestros experimentos y por los del grupo XRCE son bastante similares, y superan ampliamente a la media del resto de participantes.

Se analizan a continuación, de manera independiente, los resultados obtenidos en cada una de las ediciones. En primer lugar, se muestran en la Tabla 7-31 los resultados de evaluación correspondientes a la edición de 2010 para todos los grupos participantes que presentaron experimentos basados en aproximaciones multimodales (un total de 72 experimentos presentados por 9 grupos diferentes):

Tabla 7-31. Fusión multimedia en *ImageCLEF* 2010

Grupo	MAP	P@10	P@20
LSMF	**0,3111**	**0,5929**	**0,5479**
xrce	0,2765	0,5814	0,5193
telecom	0,2026	0,4914	0,4336
duth	0,1998	0,5200	0,4836
i2rcviu	0,1984	0,4971	0,4321
sztaki	0,1794	0,4857	0,4329
nus	0,0758	0,2671	0,2321
rgu	0,0617	0,2271	0,2129
uaic	0,0423	0,1543	0,1529
MEDIA	0,1387	0,3701	0,3293

Puede observarse que la aproximación multimedia desarrollada en base a la estrategia de fusión presentada en esta tesis, es la que mejor rendimiento ofrece de entre todos los experimentos presentados al *ImageCLEF* en la edición de 2011 de la tarea de recuperación. El mejor experimento del resto de participantes es el correspondiente al grupo *xrce* de Xerox (Clinchant et al., 2010), que utiliza diferentes operadores de agregación para combinar las listas de resultados monomodales previamente normalizadas. Sus mejores resultados son obtenidos aplicando un operador basado en la media aritmética ponderada, aunque sus resultados quedan por debajo de los aquí presentados tanto en términos de precisión global (MAP) como de precisión en los primeros resultados (P@10 y P@20). La diferencia obtenida puede considerarse relevante (11% de mejora en MAP), aunque no pueden llevarse a cabo pruebas de significancia estadística al no disponerse de los resultados detallados por consulta para los experimentos de los grupos participantes en la competición. El tercer clasificado, el grupo *telecom* (Popescu, 2010), aplica un reranking de los resultados textuales utilizando modelos de consulta extraídos desde *Flickr*, obteniendo unos resultados en la recuperación sensiblemente inferiores a los obtenidos siguiendo la estrategia LSMF (que obtiene una

mejoría en términos de MAP de más del 50%). En cuanto al resto de grupos, ninguno de ellos alcanza un 20% de precisión global, siguiendo estrategias de combinación que apenas mejoran sus propios resultados textuales, como en el caso del grupo *duth* (Arampatzis, Chatzichristofis and Zagoris, 2010) que utiliza dos tipos de métodos de fusión: uno basado en la normalización de *scores* y otro en su combinación. Otro tipo de aproximaciones observadas son las basadas en la expansión de la consulta textual en base a los resultados visuales. Es el caso del grupo *sztaki* (Daroczy, Petras and Benczur, 2010), que logra una ligerísima mejora de sus resultados textuales gracias a esta aproximación multimedia. La identificación de conceptos a partir de las imágenes tampoco ofrece un rendimiento adecuado si se observan los resultados obtenidos por el grupo *nus*. También se utilizan técnicas de fusión a nivel de características (*early fusion*) por parte del grupo *rgu* (Wang, Song and Kaliciak, 2010), desarrollando un sistema compuesto y no separable de características textuales y visuales, obteniendo unos resultados bastante lejanos a los obtenidos en esta tesis.

Se continua en la Tabla 7-32 con la revisión y el análisis de los resultados obtenidos en la edición de 2011 del *ImageCLEF* por parte de los grupos presentados a la tarea de recuperación, con el objetivo de compararlos con la propuesta de fusión multimedia descrita en esta tesis (LSMF).

Tabla 7-32. Fusión multimedia en *ImageCLEF* 2011

Grupo	MAP	P@10	P@20
xrce	0,3880	0,6320	0,5100
LSMF	**0,3400**	**0,5540**	**0,4550**
cea list	0,3075	0,5420	0,4210
duth	0,2886	0,4860	0,3870
demir	0,2432	0,4520	0,3420
dbisformat	0,2195	0,4180	0,3630
sztaki	0,2167	0,4700	0,3690
uaic20111	0,1665	0,4080	0,3090
MEDIA	0,2558	0,4542	0,3678

En este caso, y como se observa en la tabla, el mejor rendimiento es logrado por el grupo de investigación de Xerox (*xrce*), el cual sigue una técnica de fusión multimedia similar a LSMF, mediante la que combinan su propios resultados con los obtenidos por el grupo *cea_list*

(Csurka, Clinchant and Popescu, 2011). La aproximación utilizada por este grupo aplica un *reranking* a sus propios resultados textuales en base a modelos de consulta extraídos desde *Flickr*, y también hacen uso de conceptos visuales, como *outdoor/indoor*, para caracterizar las consultas en relación a si dichos conceptos aparecen o no en las imágenes (con lo que posteriormente reordenan los resultados finales). El grupo *duth* (Arampatzis, Konstantinos and Chatzichristofis, 2011), que también obtiene unos resultados aceptables (aunque por debajo de LSMF), incluye una estimación acerca de la dificultad de la consulta, cuya combinación con una combinación basada en los *scores* genera sus mejores resultados. El resto de grupos, como *demir* (Berber et al., 2011) o *sztaki* (Daroczy, Pethes and Benczur, 2011), presentan combinaciones multimedia basadas en fusión tardía con las que apenas consiguen mejoras en relación a sus aproximaciones textuales (incluso a veces el rendimiento es peor).

En resumen, queda constatado el buen funcionamiento de la estrategia de fusión multimedia propuesta en esta tesis. Los resultados mostrados sobre su rendimiento la sitúan al nivel de la mejor (o incluso superior en algunos casos) y muy por encima de la mayoría de aproximaciones observadas (más de un 70% de mejora relativa en MAP comparándolo con la media obtenida por los grupos participantes durante las ediciones de 2010 y 2011 de la tarea de recuperación multimedia de imágenes de Wikipedia). Además, y como ya se demuestra en el apartado anterior (7.3.3), la propuesta de fusión LSMF es capaz de mejorar los resultados obtenidos por la aproximación basada en texto (TBIR) gracias al aprovechamiento de la información proporcionada desde la modalidad visual de las imágenes de la colección (y de las consultas multimedia). Remarcar por último que estas mejoras son logradas a la vez que el proceso global de recuperación es considerablemente simplificado, gracias a la reducción de la colección multimedia original por parte del prefiltro textual, lo que facilita el trabajo del sistema visual (CBIR).

7.3.5 Normalización previa a la fusión multimedia

En esta sección se analiza la conveniencia de incluir una fase de normalización de los *scores* o valores de relevancia asignados a los objetos multimedia recuperados en las listas de resultados, previa a la fase de fusión multimedia tardía (*late fusion*), dentro del enfoque propuesto en esta tesis (LSMF).

Como se ha descrito en el apartado 3.6, la normalización en una tarea de fusión de resultados consiste en hacer que el rango de los scores calculados para los objetos recuperados desde diferentes sistemas de recuperación, sean comparables entre ellos (Wu, Crestani and Bi, 2006). En la experimentación mostrada hasta el momento en esta tesis, las listas de resultados procedentes del sistema de recuperación visual (CBIR) utilizan valores de relevancia o *scores* entre 0 y 1, mientras que las que llegan de los sistemas textuales (TBIR) no han sido acotados dentro de ningún rango, sino que han mantenido el *score* original obtenido desde la herramienta Lucene. Para el caso CBIR, y como se explica detalladamente en el apartado 6.3, los *scores* visuales hacen referencia a la probabilidad de que una imagen dada pertenezca a un conjunto determinado (imágenes de ejemplo), por lo que dichos *scores* estarán siempre entre 0 y 1. Por el contrario, para el caso TBIR (descrito en el apartado 6.2), los valores de similitud generados mediante la herramienta Lucene no están dentro del rango 0-1 ya que esta no normaliza el producto escalar de la misma manera que la función coseno original, al desviarse de la teoría del VSM.

Para analizar el rendimiento de la técnica de combinación LSMF cuando se añade la etapa de normalización, se evalúan los resultados en base a dos formas diferentes de normalizar los *scores* visuales. A estas aproximaciones se las ha llamado en este trabajo normalización básica y normalización "perfecta", y son descritas a continuación:

- Normalización básica. Consiste en dividir los valores de relevancia obtenidos para una determinada consulta entre el máximo obtenido para la consulta correspondiente. De este modo se consigue que el rango de valores quede delimitado entre 0 y 1, correspondiendo el 1 al score (tras normalizar) del primer resultado recuperado para cada consulta.

- Normalización "perfecta". Parte de la idea de que para que un documento sea recuperado con valor de relevancia 1, este debería ser exactamente igual a la consulta introducida (similitud máxima). Para cada consulta, se construye un documento con idéntico contenido textual que el de la consulta. Este documento "perfecto" es indexado como parte de la colección. Cuando se lanza la consulta, el documento "prefecto" creado artificialmente es recuperado en primera posición, y su *score* será considerado como el máximo posible. Por esto, este *score* máximo será utilizado para dividir todos los *scores* de la misma consulta y, de este modo,

generar una lista de resultados con valores de relevancia entre 0 y 1. El valor 1 será el asignado al documento "perfecto", el cual se eliminará de la lista de resultados ya que fue incluido artificialmente.

Se recuerda que la colección es multilingüe y que, como se vio en el apartado 7.1.3, la estrategia seguida por TBIR está basada en recuperaciones independientes sobre índices monomodales, cuyos resultados son posteriormente combinados siguiendo una aproximación de fusión tardía. Por esto, el proceso de normalización "perfecta" descrito necesitará de un documento "perfecto" para cada idioma, a partir del que se conseguirá una lista de resultados normalizada también para cada idioma. Finalmente, y para completar el proceso, habrá que combinar las listas de resultados normalizadas monolingües mediante un algoritmo de fusión tardía para obtener la lista definitiva multimedia fusionada y normalizada. El algoritmo de fusión utilizado será *MaxMerge* (descrito en 6.4), utilizado para combinar los resultados monolingües en los experimentos textuales originales (sin normalización), el cual selecciona las imágenes recuperadas con mayor *score* independientemente del idioma utilizado para su recuperación.

Los experimentos relacionados con el análisis de la normalización de *scores* previa a la fusión multimedia se llevan a cabo dentro del enfoque que utiliza LSMF, por lo que la comparación se hace con los resultados mostrados en el apartado 7.3.3.

Se muestran en primer lugar los resultados correspondientes al caso de utilizar la normalización básica.

Tabla 7-33. Normalización básica (o no) dentro de LSMF

		MAP		**P@5**		**P@10**		**P@20**	
Normalización		**no**	**básica**	**no**	**básica**	**no**	**básica**	**no**	**básica**
TBIR		0,3044		0,5600		0,5060		0,4040	
CBIR prefiltrado		0,0618		0,0880		0,0880		0,0910	
Product		0,3400	0,3400	0,6600	0,6600	0,5540	0,5540	0,4550	0,4550
OWA	min	0,1777	**0,3039**	0,3520	**0,5920**	0,3160	**0,5040**	0,2660	**0,4170**
	orness01	0,3103	**0,3210**	0,6360	**0,6400**	0,5280	0,5220	0,4290	0,4250
	orness02	0,3369	0,3304	0,6600	0,6480	0,5660	0,5400	0,4450	0,4380
	orness03	0,3371	0,3358	0,6440	**0,6680**	0,5540	0,5320	0,4480	0,4470
	orness04	0,3319	**0,3348**	0,6200	**0,6560**	0,5420	**0,5540**	0,4330	**0,4500**
	avg	0,3233	0,3224	0,6000	**0,6360**	0,5240	**0,5560**	0,4350	**0,4460**
	orness06	0,3174	0,3054	0,5840	**0,6360**	0,5200	**0,5560**	0,4240	**0,4310**
	orness07	0,3108	0,2719	0,5800	**0,6000**	0,5080	**0,5200**	0,4130	**0,4210**
	orness08	0,3049	0,2326	0,5720	**0,5760**	0,5060	0,4700	0,4080	0,3680
	orness09	0,2994	0,1977	0,5720	0,5240	0,5000	0,4160	0,4020	0,3120
	max	0,2933	0,1731	0,5600	0,4720	0,5000	0,3620	0,3980	0,2690
FilterN	N=1000	0,2395	0,2395	0,5720	0,5720	0,5000	0,5000	0,4000	0,4000
	N=1500	0,2900	0,2900	0,5680	0,5680	0,5040	0,5040	0,4030	0,4030
	N=2000	0,3065	0,3065	0,5640	0,5640	0,5100	0,5100	0,4040	0,4040
	N=2500	0,3066	0,3066	0,5640	0,5640	0,5100	0,5100	0,4040	0,4040
	N=5000	0,3056	0,3056	0,5640	0,5640	0,5080	0,5080	0,4040	0,4040
	N=10000	0,3048	0,3048	0,5600	0,5600	0,5060	0,5060	0,4040	0,4040
Enrich	x1	0,3079	0,2921	0,5640	0,5080	0,5080	0,4760	0,4050	0,3960
	x1.25	0,3074	0,2818	0,5560	0,4920	0,5080	0,4680	0,4060	0,3940
	x1.5	0,3079	0,2669	0,5600	0,4840	0,5040	0,4600	0,4070	0,3920
	x2	0,3043	0,2606	0,5520	0,4360	0,5000	0,4460	0,4050	0,3850
	x5	0,2766	0,2228	0,4720	0,2320	0,4500	0,3500	0,3900	0,3490
MaxMerge		0,2933	0,1731	0,5600	0,4720	0,5000	0,3620	0,3980	0,2690

Lo primero a destacar de los resultados observados es el hecho de que el rendimiento obtenido por la fusión cuando se utilizan los algoritmos *Product* y *FilterN* no se ve afectado por la inclusión o no del proceso de normalización de *scores*. La explicación en el caso del algoritmo *Product*, es que la fusión se lleva a cabo mediante la multiplicación de los scores correspondientes a las listas de entrada, por lo que no influye el hecho de normalizar los valores de dichas listas para la ordenación final de los elementos recuperados en la lista fusionada. Para el caso del algoritmo *FilterN*, la estrategia de fusión consiste en eliminar de la lista de resultados textuales (TBIR) aquellos que no encuentren entre los primeros N de la lista visual (CBIR), esto es, está basada en las posiciones ocupadas por los elementos recuperados y no en los *scores* o valores de relevancia de estos, por lo que la normalización o no de dichos valores no afectará al proceso de fusión.

Por otro lado, también puede observarse en la Tabla 7-33, cómo los resultados obtenidos por el algoritmo *Enrich* no mejoran en ningún caso, para ninguna de las medidas de evaluación analizadas (MAP, P@5, P@10 y P@20) y con ninguna de las configuración de los parámetros del algoritmo, los resultados obtenidos cuando se utilizan los *scores* originales de la lista textual (TBIR). Esto parece indicar que la influencia producida por los resultados visuales (CBIR) no es del todo positiva, ya que cuando se utilizan rangos similares para los valores de relevancia de ambas listas, los resultados fusionados definitivos empeoran. Para el caso del algoritmo *MaxMerge*, también se da la situación de que los resultados de la fusión empeoran cuando se lleva a cabo el proceso de normalización de la lista visual. En este caso el empeoramiento es todavía peor que en el caso del *Enrich*, ya que lo que se hace es seleccionar el valor de relevancia máximo entre los obtenidos por una determinada imagen en ambas listas de resultados (TBIR y CBIR). Por este motivo, y ya que los mejores resultados se obtienen desde la aproximación TBIR, cuando se iguala el rango de valores para los *scores* de ambas listas los resultados fusionados son peores al darle demasiada importancia a los valores de similitud visuales.

El único algoritmo de fusión para el que se obtienen, solo en algunos casos, mejores resultados cuando se incluye el paso de normalización que cuando no, es el OWA. Para la configuración más restrictiva de este algoritmo (*orness* = 0.0, que funciona como un AND), es cuando los resultados de fusión mejoran claramente cuando se aplica la normalización de TBIR. Para este caso, la mejora es considerable para todas las medidas de evaluación, ya que cuando no se normaliza se seleccionan en la gran mayoría de los casos los valores correspondientes a la lista de resultados CBIR, ya que esta está expresada dentro de un rango de valores mucho menor que la lista TBIR. Para la configuración que mejor funcionaba este algoritmo ($orness = 0{,}2$, tal y como se vio en la tabla Tabla 7-25), los resultados de la fusión son bastante parecidos, aunque ligeramente inferiores, cuando se añade la fase de normalización para los *scores* textuales. Cabe destacar el valor de P@5 (0,6680) obtenido con este algoritmo de fusión cuando se utiliza un valor de $orness = 0{,}3$ y se incluye el proceso de normalización, que es el mayor valor obtenido hasta el momento, por encima del correspondiente al del algoritmo de fusión *Product*. Con valores de *orness* entre 0.4 y 0.7 también se mejoran ligeramente los resultados obtenidos cuando no se normaliza para las medidas de evaluación P@10 y P@20.

La conclusión extraída de este conjunto de experimentos destinados a analizar la conveniencia de la normalización de la lista de resultados textuales (TBIR) antes de aplicar los algoritmos de fusión de las listas de resultados monomodales dentro de la estrategia de combinación multimedia LSMF, es que dicha normalización no mejora en la mayoría de los casos ni para la mayoría de los algoritmos. Se ha visto cómo a los algoritmos *FilterN* y *Product* la normalización no le afecta en absoluto (por su propia definición), mientras que para *Enrich* y *MaxMerge* la normalización de la lista TBIR empeora los resultados en todos los casos. Solo para algunas de las configuraciones del algoritmo de fusión OWA, los resultados de la fusión son mejores cuando se aplica la normalización que cuando no.

Se analizan ahora los resultados de la combinación LSMF cuando se hace uso de la otra aproximación de normalización descrita, la normalización "perfecta". La siguiente tabla compara los resultados con aquellos desarrollados en el apartado 7.3.3, donde no se aplica ningún tipo de normalización sino que se utilizan los valores de relevancia originales de cada uno de los sistemas de recuperación monomodales. No se incluyen en esta tabla los resultados correspondientes a los algoritmos *Product* y *FilterN*, para los que ya se vio que la normalización no afecta a sus resultados debido a la propia definición teórica de dichos algoritmos.

Tabla 7-34. Normalización "perfecta" (o no) dentro de LSMF

		MAP		P@5		P@10		P@20	
Normalización		**no**	**perfecta**	**no**	**perfecta**	**no**	**perfecta**	**no**	**perfecta**
TBIR		0,3044		0,5600		0,5060		0,4040	
CBIR prefiltrado		0,0618		0,0880		0,0880		0,0910	
OWA	min	0,1777	**0,3062**	0,3520	**0,5440**	0,3160	**0,5120**	0,2660	**0,4050**
	orness01	0,3103	**0,3218**	0,6360	0,6000	0,5280	0,5180	0,4290	0,4210
	orness02	0,3369	0,3257	0,6600	0,6360	0,5660	0,5260	0,4450	0,4300
	orness03	0,3371	0,3188	0,6440	**0,6680**	0,5540	0,5400	0,4480	0,4330
	orness04	0,3319	0,2996	0,6200	**0,6760**	0,5420	0,5360	0,4330	0,4210
	avg	0,3233	0,2692	0,6000	**0,6360**	0,5240	**0,5320**	0,4350	0,3990
	orness06	0,3174	0,2266	0,5840	0,5480	0,5200	0,4620	0,4240	0,3400
	orness07	0,3108	0,1726	0,5800	0,4360	0,5080	0,3660	0,4130	0,2710
	orness08	0,3049	0,1290	0,5720	0,3080	0,5060	0,2500	0,4080	0,2050
	orness09	0,2994	0,1007	0,5720	0,2040	0,5000	0,1780	0,4020	0,1550
	max	0,2933	0,0823	0,5600	0,1680	0,5000	0,1440	0,3980	0,1260
Enrich	x1	0,3079	0,2602	0,5640	0,4240	0,5080	0,4260	0,4050	0,3650
	x1.25	0,3074	0,2513	0,5560	0,3840	0,5080	0,4140	0,4060	0,3620
	x1.5	0,3079	0,2441	0,5600	0,3680	0,5040	0,4140	0,4070	0,3580
	x2	0,3043	0,2319	0,5520	0,3120	0,5000	0,3740	0,4050	0,3460
	x5	0,2766	0,1914	0,4720	0,1280	0,4500	0,2200	0,3900	0,2870
MaxMerge		0,2933	0,0823	0,5600	0,1680	0,5000	0,1440	0,3980	0,1260

Los resultados de fusión que se observan tras aplicar este tipo de normalización siguen un comportamiento similar a los obtenidos con la otra aproximación de normalización analizada anteriormente. Para los algoritmos *Enrich* y *MaxMerge*, los resultados de la fusión no mejoran a los obtenidos a partir de los monomodales (TBIR y CBIR) sin normalizar. Los valores de precisión, tanto a nivel general (MAP) como a nivel de primeras posiciones (P@5, P@10 y P@20), obtenidos por estos algoritmos tras esta normalización resultan inferiores a los correspondientes a la normalización básica. La explicación a esto se corresponde con la dada en el análisis de resultados anterior, donde se comentaba que la influencia de los resultados visuales no era positiva, por lo que al disminuir el rango de valores de la lista textual, los resultados fusionados empeora. Y en este caso, empeoran incluso más, ya que la normalización llevada a cabo reduce todavía más los valores o *scores* textuales, ya que divide estos entre un valor más alto (el correspondiente al documento "perfecto" artificial, como se explica en la definición de este tipo de normalización).

Para el caso del algoritmo *OWA* los resultados son más bajos cuando se incluye la normalización que cuando no. Como sucedía con el otro tipo de normalización, para ciertas

configuraciones del algoritmo los resultados mejoran, aunque en este caso se dan menos situaciones de este tipo. Ahora ningún valor de *orness* mayor que 0.5 es capaz de mejorar los resultados para ninguna de las medidas de evaluación analizadas. Solo para el caso del $orness = 0$ se mejoran todos los resultados, y ya se explicó que el motivo de esto es que esta configuración del algoritmo OWA cuando no se normaliza, selecciona en la práctica totalidad de los casos los valores de relevancia visuales, ya que por lo general resultan ser menores. Por esto se obtiene un rendimiento tan bajo tras la fusión (similar al de la recuperación visual monomodal), y este es mejorado cuando se aplica cualquier tipo de normalización sobre la lista de resultados textuales. Por último, cabe destacar el valor de P@5 obtenido cunado se configura el algoritmo OWA con un $orness = 0{,}4$, para la que se obtiene el mejor valor de evaluación para esta medida de precisión temprana ($67{,}60\%$).

En la siguiente figura pueden verse gráficamente los resultados obtenidos por las dos propuestas de normalización analizadas, comparándolas entre ellas y también con la propuesta mostrada en el apartado 7.3.3 que no aplica ningún tipo de normalización. Se muestran en la gráfica los resultados obtenidos para cada uno de los algoritmos de fusión multimedia utilizados dentro de la propuesta LSMF, seleccionando en cada caso la mejor configuración (la decidida en el apartado 7.3.3). Se muestran únicamente los resultados para los algoritmos *OWA*, *Enrich y* MaxMerge ya que, como se ha visto anteriormente, el rendimiento de los algoritmos *Product* y *FilterN* es independiente de la presencia o no de la fase de normalización.

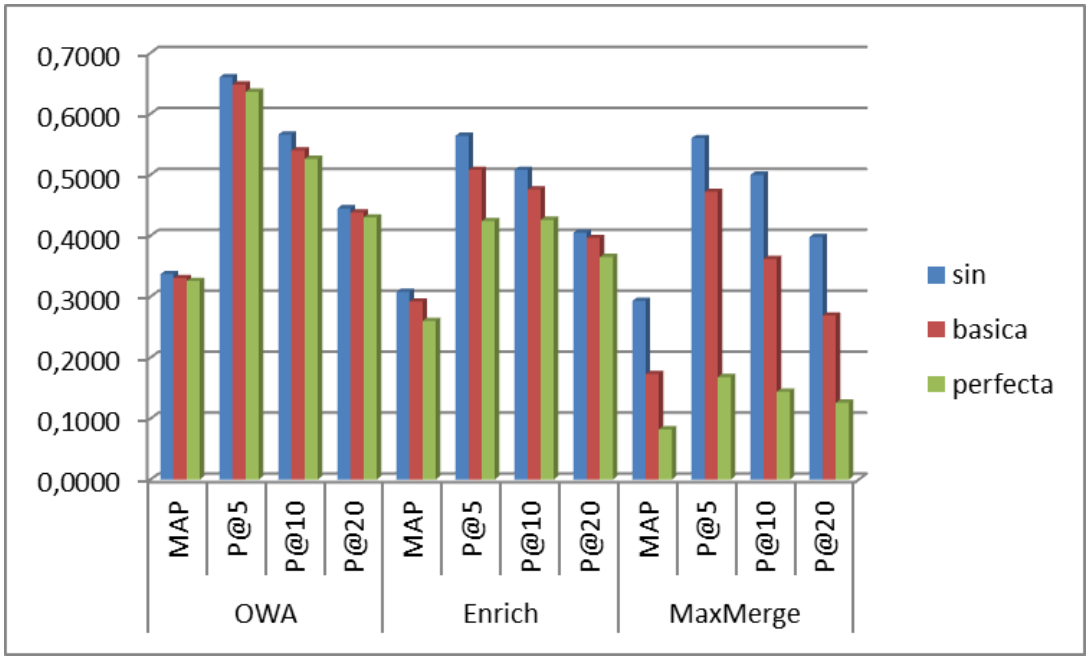

Figura 7.8. Comparativa entre normalización básica, "perfecta", o ninguna (sin)

Analizando los resultados se observa que el rendimiento de los diferentes experimentos es bastante parecido, si bien la aproximación que mejores resultados ofrece es cuando se trabaja sin la normalización previa a la fase de prefiltrado. Como se ha visto a lo largo de este apartado, sí que existen algunas configuraciones concretas que mejoran algunos valores de evaluación cuando se normaliza en comparación con cuando no, pero en general lo que se extrae de este conjunto de experimentos es la no necesidad de llevar a cabo la fase de normalización de la lista de resultados textual. Solo para el caso del algoritmo de fusión *MaxMerge*, la diferencia de rendimientos es clara y a favor de la no normalización de *scores* textuales, debido a la propia definición del algoritmo que selecciona para la lista fusionada final el *score* más alto de las listas combinadas.

Como se ha visto en el apartado correspondiente al estado del arte (apartado 3.6), el rendimiento de un sistema que utiliza valores de relevancia o *score* normalizados para llevar a cabo el proceso de fusión (multimedia o no), dependerá en gran medida de la propia definición de dichos scores en cada experimento (Zhou, Depeursinge and Müller, 2010). Se ha visto aquí como para el caso concreto de la fusión multimedia entre listas de resultados basadas en texto y en imagen para una tarea de recuperación, la inclusión de una fase de normalización de la lista textual (la visual ya está originalmente normalizada entre 0 y 1) no

resulta necesaria para la mayor parte de los experimentos analizados. Es más, las dos aproximaciones de normalización aplicadas (básica y "perfecta", descritas anteriormente) obtienen un rendimiento más bajo que la aproximación que no aplica normalización alguna.

7.3.6 LSMF con CBIR por tipos de descriptores visuales

En este apartado se llevan a cabo una serie de experimentos dedicados a fusionar los resultados CBIR (tras la aplicación del prefiltro textual) obtenidos en base a diferentes grupos de descriptores visuales de manera independiente, con los resultados textuales (TBIR). La estrategia de combinación LSMF aplica inicialmente el prefiltro textual, con lo que se trabajará sobre la subcolección obtenida. Se hace uso de la lista de imágenes resultado obtenida mediante la recuperación textual, como puede verse en la Figura 7.9, así como de las diferentes listas de resultados por grupos de descriptores visuales descritas en el apartado 7.2.2. La idea es comparar el funcionamiento de la estrategia de fusión LSMF cuando se fusionan las modalidades textual y visual en diferentes momentos: 1) tras la recuperación monomodal global (experimento apartado 7.3.3), y 2) CBIR independiente para cada conjunto de descriptores visuales definidos, y posterior combinación de resultados.

La experimentación desarrollada en este apartado se lleva a cabo a partir de las listas de resultados obtenidas tras la recuperación visual (CBIR) basada en grupos de descriptores visuales (apartado 7.2.2, Figura 7.1), y de los resultados textuales obtenidos desde el sistema TBIR para el conjunto de consultas de evaluación de la edición de 2011 del *ImageCLEF*. Se muestra a continuación la diferencia entre las dos alternativas a comparar:

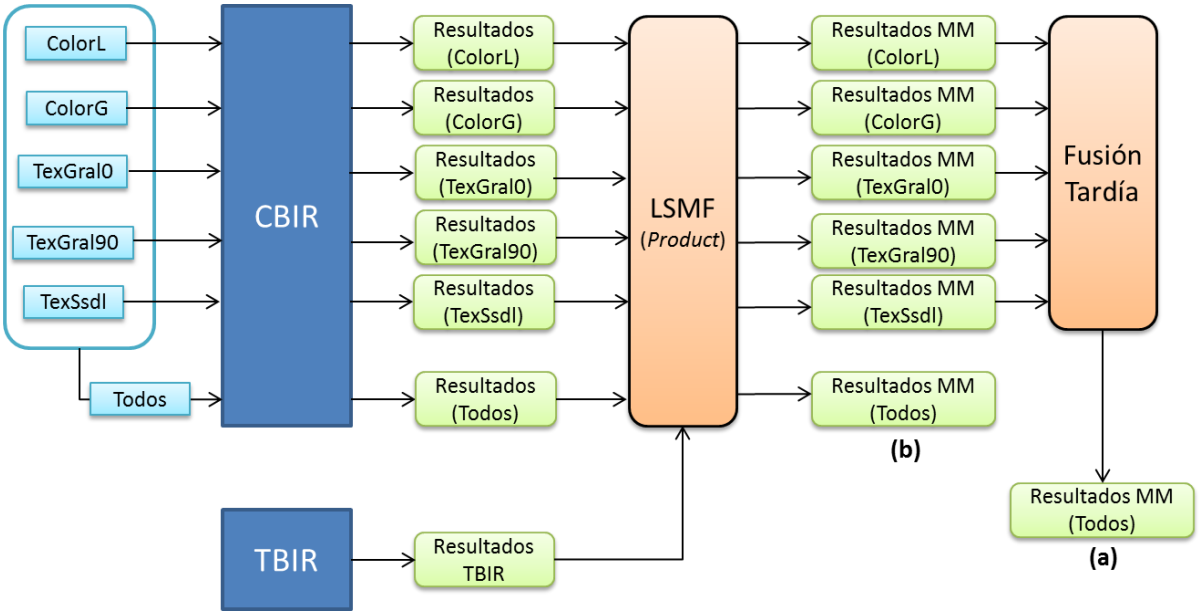

Figura 7.9. Alternativas LSMF con descriptores visuales

Tal y como muestra la figura, cada una de las listas de resultados visuales obtenidas gracias al sistema CBIR en base a los distintos grupos de descriptores definidos (ColorL, ColorG, TexGral0, TexGral90, TexSsdl), se combina de manera independiente con la lista de resultados textuales obtenida desde el sistema TBIR. Para esta combinación se sigue la aproximación LSMF con el algoritmo de fusión tardía *Product*, que fue el que mejores resultados ofreció según el análisis del apartado 7.3.3. Los resultados de estas combinaciones son los que se muestran en la Tabla 7-35. En la misma se incluyen también los resultados correspondientes a la fusión multimedia LSMF que utiliza los resultados CBIR basados en todos los descriptores visuales de manera conjunta.

Tabla 7-35. Resultados LSMF por grupos de descriptores visuales

Resultados MM	**MAP**	**P@5**	**P@10**	**P@20**
LSMF_Product(ColorG)	0,2984	0,5800	0,4800	0,4000
LSMF_Product(ColorL)	0,3308	0,6040	0,5420	0,4390
LSMF_Product(TexGral0)	0,2936	0,5752	0,4940	0,3910
LSMF_Product(TexGral90)	0,2951	0,5880	0,5040	0,3870
LSMF_Product(TexSsdl0)	0,2847	0,5560	0,4900	0,3800
LSMF_Product(Todos)	0,3404	0,6280	0,5480	0,4530

Los resultados obtenidos tras esta primera fase de la experimentación muestran un comportamiento similar a los correspondientes a los puramente visuales comentados en el apartado 7.2.2, también basados en grupos de descriptores visuales independientes. Se observa que ninguna de las recuperaciones basadas en grupos de descriptores mejora el rendimiento de utilizar todos esos descriptores de manera conjunta en la fase de recuperación visual. El conjunto de características visuales que mejor funciona es el correspondiente al color general (ColorG), que obtiene un valor de MAP bastante cercano al obtenido por el experimento que utiliza todas las características de forma conjunta.

El último paso para completar el experimento planteado consistirá en combinar entre sí las cinco listas obtenidas tras aplicar LSMF entre TBIR y CBIR (para cada uno de los grupos de descriptores visuales), como se muestra en la Figura 7.9. Para llevar a cabo esta combinación se hace uso de los algoritmos de fusión tardía *Product* y *MaxMerge*, que permiten combinar más de dos listas de resultados. La Tabla 7-36 muestra los valores de evaluación obtenidos por estos dos algoritmos, incluyendo también los resultados de LSMF que utiliza los resultados CBIR generales.

Tabla 7-36. Tras combinación de las listas fusionadas por grupos de descriptores visuales

	MAP	**P@5**	**P@10**	**P@20**
MaxMerge(ResultadosMM)	0,3028	0,5720	0,5100	0,3920
Product(ResultadosMM)	0,3098	0,6160	0,5280	0,4210
LSMF_Product(Todos)	0,3404	0,6280	0,5480	0,4530

De entre los dos algoritmos de fusión tardía evaluados para la realización de este último paso, es ligeramente mejor el que utiliza el algoritmo *Product* para todas las medidas de evaluación analizadas. Sin embargo, en ninguno de los dos casos se consigue mejorar los resultados de la recuperación multimedia basada en el uso de todos los descriptores visuales de forma conjunta (y no por grupos).

En definitiva, la experimentación llevada a cabo con el fin de analizar la conveniencia de tratar las características visuales de manera independiente, no recomienda dicha aproximación. Los resultados de recuperación empeoran en todos los casos (en comparación con su uso de forma conjunta) y, además, se añaden varias fases de fusión de resultados

adicionales (una para cada grupo de descriptores visuales diferenciado, y otra para combinar lo obtenido desde cada grupo). Por lo tanto, debido a los resultados de precisión obtenidos y al aumento de la complejidad, no se considera recomendable la aproximación de fusión multimedia basada en grupos de descriptores visuales independientes (al menos trabajando dentro de la estrategia LSMF propuesta en esta tesis).

7.4 Análisis en función de la complejidad y la carga visual de las consultas

En esta sección se lleva a cabo un análisis del rendimiento de la estrategia de fusión propuesta (LSMF) en función de la dificultad y la carga visual o "visualidad" de las consultas lanzadas (Popescu, Tsikrika and Kludas, 2010). Para clasificar en función de la dificultad se utilizan los promedios de los valores de precisión media (AP, *Average Precision*) obtenidos para cada consulta en todos los experimentos presentados por los grupos participantes, del siguiente modo:

Tabla 7-37. Umbrales para clasificar dificultad de consultas

Dificultad	MAP obtenido
Fácil	$> 0{,}3$
Media	$0{,}2 \geq MAP \leq 0{,}3$
Alta	$0{,}1 \geq MAP \leq 0{,}2$
Muy alta	$< 0{,}1$

De las 70 consultas multimedia propuestas en 2010, 53 de ellas se clasifican dentro de las categorías de dificultad alta o muy alta, ya que se persigue, intencionadamente, disponer de consultas altamente semánticas, ofreciendo así un interesante desafío para las estrategias de recuperación. Cuatro de ellas obtienen un $MAP < 0{,}05$ ("*woman in red dress*", "*building site*", "*horseman*", "*people laughing*") y son consideradas irresolubles. De todo el conjunto de consultas, solo cuatro de ellas son consideradas como fáciles ("*satellite image*", "*portrait of Jintao Hu*", "*ferrari red*", "*postage stamp*").

La característica de visualidad se define en función de los experimentos puramente textuales (TBIR) y de los que utilizan tanto recursos textuales como visuales (MMIR, *multimedia image retrieval*). Los experimentos TBIR superan a los MMIR en 37 de las 70 consultas, y

son superados por estos en 31. Esto significa que en menos de la mitad de las consultas las aproximaciones multimedia aprovechan la multimodalidad para mejorar la recuperación basada únicamente en texto. En base a la diferencia en cuanto a los resultados obtenidos en las distintas consultas por los sistemas TBIR y MMIR, se define la propiedad de visualidad. Una consulta es "Visual" cuando $diff(MAP \geq 0{,}01)$ a favor de la aproximación TBIR. De igual modo una consulta es considerada "Textual" cuando la misma diferencia se da a favor de las soluciones MMIR. Según esta definición, para las consultas multimedia de la edición 2010, 26 de ellas son consideradas textuales y 24 visuales. Las 20 consultas restantes se clasifican como neutras al no observarse diferencias significativas entre la aproximación textual y la multimedia.

En cuanto a las 50 consultas multimedia propuestas para la tarea de recuperación de imágenes de la edición de 2011, y siguiendo los criterios anteriormente descritos, 21 de ellas se clasifican como difíciles o muy difíciles, debido de nuevo a la intención de crear, como en la edición de 2010, consultas altamente semánticas. En relación a la visualidad de este conjunto de consultas, 38 de ellas son consideradas visuales y 7 textuales. Las 5 restantes se clasifican como neutras. En comparación con la edición de 2010, existen más consultas consideradas visuales, debido al mayor número de imágenes de ejemplo proporcionadas dentro de la parte visual de las consultas multimedia propuestas, lo que permite mejorar los resultados de CBIR.

Tabla 7-38. Clasificación de Dificultad y Visualidad de las 50 consultas multimedia en *ImageCLEF* 2011

Consulta	Parte Textual (inglés)	Dificultad	Visualidad
71	colored Volkswagen beetles	Fácil	Visual
72	skeleton of dinosaur	Fácil	Visual
73	graffiti street art on walls	Media	Visual
74	white ballet dress	Difícil	Textual
75	flock of sheep	Fácil	Neutra
76	playing cards	Fácil	Neutra
77	cola bottles or cans	Fácil	Textual
78	kissing couple	Difícil	Textual
79	heart shaped	Media	Textual
80	wolf close up	Difícil	Visual
81	golf player on green	Difícil	Visual
82	model train scenery	Muy difícil	Visual
83	red or black mini cooper	Media	Neutra
84	Sagrada Familia in Barcelona	Fácil	Visual

85	Beijing bird nest	Fácil	Visual
86	KISS live	Fácil	Visual
87	boxing match	Muy difícil	Visual
88	portrait of Segolene Royal	Fácil	Visual
89	Elvis Presley	Fácil	Visual
90	gondola in Venice	Difícil	Visual
91	freestyle jumps with bmx or motor bike	Media	Visual
92	air race	Media	Visual
93	cable car	Difícil	Visual
94	roller coaster wide shot	Fácil	Visual
95	photo of real butterflies	Difícil	Visual
96	shake hands	Fácil	Textual
97	round cakes	Media	Visual
98	illustrations of Alice's adventures in Wonderland	Fácil	Neutra
99	drawings of skeletons	Difícil	Visual
100	brown bear	Fácil	Visual
101	fountain with jet of water in daylight	Difícil	Visual
102	black cat	Muy difícil	Visual
103	dragon relief or sculpture	Difícil	Textual
104	portrait of Che Guevara	Media	Visual
105	chinese characters	Difícil	Visual
106	family tree	Media	Visual
107	sunflower close up	Fácil	Visual
108	carnival in Rio	Media	Visual
109	snowshoe hiking	Fácil	Visual
110	male color portrait	Muy difícil	Visual
111	two euro coins	Media	Visual
112	yellow flames	Difícil	Visual
113	map of Europe	Fácil	Neutra
114	diver underwater	Media	Visual
115	flying bird	Muy difícil	Visual
116	houses in mountains	Muy difícil	Visual
117	red roses	Difícil	Visual
118	flag of UK	Muy difícil	Textual
119	satellite image of desert	Difícil	Visual
120	bar codes	Media	Visual

7.4.1 Análisis del Prefiltro textual

Para evaluar el funcionamiento del prefiltro textual semántico utilizado en la primera fase de la estrategia de fusión multimedia LSMF propuesta en esta tesis, se analizan los resultados de dicho prefiltro para los diferentes tipos de consultas multimedia. Las tablas que se muestran a continuación incluyen los resultados obtenidos en base a las clasificaciones de consultas mencionadas (según su dificultad y su visualidad). En ellas se muestra, para cada tipo de

consulta multimedia, el valor de cobertura alcanzado por el conjunto de resultados obtenidos mediantes el prefiltro textual, junto con el número de imágenes relevantes recuperadas en cada caso (num_ret) y el número total de relevantes existente en la colección (num_rel). También se indican los valores de precisión a bajo nivel (*early precision*) logrados por cada experimento (P@5 y P@20).

Tabla 7-39. Resultados Prefiltro según Dificultad

Dificultad	Cobertura	num_rel	num_ret	P@5	P@20
Fácil	89,09%	909	802	0,7647	0,5676
Media	87,25%	443	381	0,5333	0,4
Difícil	82,10%	1155	920	0,5143	0,3143
Muy difícil	83,95%	933	751	0,2	0,1929

Tabla 7-40. Resultados Prefiltro según Visualidad

Visualidad	Cobertura	num_rel	num_ret	P@5	P@20
Textual	82,29%	270	215	0,5714	0,3571
Visual	85,77%	2791	2277	0,5316	0,3766
Neutral	92,64%	379	362	0,76	0,67

Los datos muestran cómo los valores de cobertura obtenidos son mejores cuanto más fáciles son las consultas (según la clasificación): las fáciles casi alcanzan el 90%, mientras que las difíciles o muy difíciles se quedan alrededor del 83%. En cualquiera de los casos, lo más interesante es resaltar la alta cobertura obtenida de media para todas las consultas, ya que esto permitirá al sistema visual de recuperación de imágenes (CBIR) trabajar, en la segunda fase de la estrategia LSMF, sobre una subcolección reducida de imágenes de calidad (recordar que la simplificación obtenida, cercana al 98%, permitirá la escalabilidad de la tarea y simplificará enormemente las fases de recuperación visual y fusión multimedia). De igual modo, los valores de precisión obtenidos son mejores para las consultas más sencillas, tanto en P@5 como en P@20.

Analizando una a una las 50 consultas multimedia, se observa que casi la mitad de ellas (22) alcanzan una cobertura casi total ($> 90\%$) en la subcolección prefiltrada. La siguiente tabla muestra este conjunto de 22 consultas indicando el valor de cobertura alcanzado, el número de imágenes relevantes y recuperadas en cada caso, y la P@5.

Tabla 7-41. Consultas con cobertura > 90%

Consulta	Dificultad	Cobertura	num_rel	num_ret	P@5
98	Easy	100 %	21	21	1.0
113	Easy	97.75 %	267	261	1.0
86	Easy	90.91 %	11	10	1.0
108	Medium	91.89 %	37	34	1.0
100	Easy	95.65 %	46	44	0.8
88	Easy	90 %	10	9	0.8
119	Hard	97.85 %	93	91	0.8
85	Easy	100 %	12	12	0.6
84	Easy	100 %	7	7	0.6
71	Easy	98 %	50	49	0.6
92	Medium	100 %	12	12	0.6
73	Medium	93.68 %	95	89	0.6
90	Hard	95.16 %	62	59	0.6
111	Medium	100 %	58	58	0.4
104	Medium	92.3 %	13	12	0.4
117	Hard	100 %	27	27	0.4
101	Hard	90.07 %	141	127	0.4
102	Very Hard	100 %	20	20	0.4
120	Medium	92.86 %	14	13	0.2
83	Medium	90 %	10	9	0.0
80	Hard	92 %	25	23	0.0
118	Very hard	91.67 %	12	11	0.0

Se observa que para estas consultas no solo se obtiene un valor de cobertura alto gracias al prefiltro textual, sino que también los valores de precisión en las cinco primeras imágenes recuperadas (P@5) son bastante buenos. Puede verse cómo las consultadas clasificadas como fáciles se encuentran en las primeras posiciones de la tabla (mayores P@5), mientras que aquellas que son consideradas difíciles o muy difíciles aparecen en la segunda mitad de la tabla (peores P@5).

El resto de consultas (no incluidas en la tabla) también obtienen buenos valores de cobertura, entre 70% y 90%, exceptuando la consulta 112 para la que se mantienen, tras el prefiltro, 54 imágenes relevantes de un total de 92 (58.7 %). La parte textual de esta consulta multimedia es *"yellow flames"*, pero las anotaciones asociadas a las imágenes que el prefiltro deja fuera utilizan otro tipo de términos para describir el mismo concepto: *"pyrotechnic", "candles"*,

"fire" or *"lighter"*. Se trata de problemas derivados del *semantic gap* textual, descrito en el apartado 2.2 de esta tesis. Este tipo de errores podría solventarse con la aplicación de técnicas de expansión de la consulta, pero se prefiere no hacerlo debido al peligro comprobado de la inclusión de demasiado ruido. Otro ejemplo de este tipo de errores entre las consultas que alcanzan una cobertura entre el 70 y el 80% puede verse en la consulta 106 (*"family tree"*), para la que no se filtran correctamente 22 imágenes de un total de 76. El motivo es similar: algunas de las imágenes relevantes perdidas están anotadas con términos como *"hierachy"* o *"genealogy"*. Como parte de una tarea de evaluación de este tipo también pueden darse errores de otra naturaleza: aquellos casos en los que una imagen es considerada relevante, pero en realidad no lo es. Esto es causa de los efectos secundarios de las técnicas de *pooling* al construir los juicios de relevancia.

7.4.2 Análisis de la fusión multimedia semántica tardía (LSMF)

Se incluye en esta sección un análisis de los resultados referentes a la aproximación textual (TBIR), y a la multimedia basada en la estrategia de fusión propuesta en esta tesis (LSMF).

La siguiente tabla muestra los valores de precisión obtenidos por las aproximaciones puramente textual (TBIR) y multimodal (LSMF) para cada uno de los tipos de consultas según su dificultad. El objetivo es comparar el rendimiento de cada una de las aproximaciones para los distintos niveles de dificultad, y analizar la mejora lograda en cada caso.

Tabla 7-42. Mejora LSMF según clasificación de Dificultad

		MAP	**P@5**	**P@10**	**P@20**
TBIR	**Fácil**	0,5140	0,7647	0,6647	0,5676
	Media	0,3047	0,5333	0,5500	0,4000
	Difícil	0,1571	0,5143	0,4071	0,3143
	Muy difícil	0,0897	0,2000	0,2429	0,1929
LSMF	**Fácil**	0,5352	0,8000	0,7118	0,5971
	Media	0,3447	0,7000	0,5583	0,4292
	Difícil	0,2080	0,6143	0,4857	0,3893
	Muy difícil	0,1218	0,3429	0,3000	0,2857
Mejora	**Fácil**	2,12%	3,53%	4,71%	2,94%
	Media	4,01%	16,67%	0,83%	2,92%
	Difícil	5,09%	10,00%	7,86%	7,50%
	Muy difícil	3,20%	14,29%	5,71%	9,29%

Lo primero destacable de los resultados de evaluación obtenidos es el hecho de que para todos los tipos de consultas (fáciles, medias, difíciles y muy difíciles) y para todas las medidas utilizadas (MAP, P@5, P@10 y P@20), los resultados tras la fusión multimedia planteada (LSMF) mejoran los correspondientes a la recuperación textual (TBIR). Estas mejoras llegan en algunos a niveles relativos de casi el 17%. La mejora observada confirma el buen comportamiento de la estrategia de fusión LSMF, que mejora a TBIR, y además simplifica enormemente el proceso global de recuperación.

Otro aspecto a destacar de los resultados es el hecho de que las mejoras relativas más significantes son conseguidas para las consultas multimedia de tipo difícil o muy difícil, con más de un 7,5% de mejora relativa media. Esto es debido a que las consultas de tipo medio, o sobre todo las de tipo fácil, ya obtienen buenos resultados en base a la aproximación textual y, por lo tanto, la mejora obtenida gracias a la fusión multimedia no es tan notable. Respecto a las medidas de evaluación utilizadas, la que obtiene mayor mejora cuando se aplica la fusión es la correspondiente la P@5 (más de 11% de mejora relativa media). Este dato indica que la aportación de los resultados visuales dentro de la estrategia de fusión incide especialmente en las imágenes recuperadas en las primeras posiciones de la lista de resultados.

Para el caso de la clasificación basada en la visualidad (o carga visual) de las distintas consultas multimedia de evaluación, se incluye la siguiente tabla que muestra los resultados obtenidos en base a la recuperación textual (TBIR) y a la estrategia de fusión multimedia propuesta (LSMF). También se incluyen los valores correspondientes a la mejora relativa obtenida gracias a dicha fusión.

Tabla 7-43. Mejora LSMF según clasificación de Visualidad

		MAP	P@5	P@10	P@20
TBIR	**Textual**	0,2503	0,5714	0,5000	0,3571
	Visual	0,2899	0,5316	0,4711	0,3776
	Neutra	0,4908	0,7600	0,7800	0,6700
LSMF	**Textual**	0,2467	0,5714	0,4714	0,3643
	Visual	0,3362	0,6579	0,5368	0,4447
	Neutra	0,4993	0,8000	0,8000	0,6600
Mejora	**Textual**	-0,36%	0,00%	-2,86%	0,71%
	Visual	4,63%	12,63%	6,58%	6,71%
	Neutra	0,85%	4,00%	2,00%	1,00%

Analizando los resultados mostrados en la tabla, se observa primeramente que tanto para las consultas de tipo visual como para las de tipo neutro se obtiene una mejora relativa de la aproximación multimedia (LSMF) en comparación con la textual (TBIR). Solo para el caso de las consultas de tipo textual, no se obtiene, por lo general, una mejora al fusionar. Además, las mejoras más importantes son las correspondientes a las consultas de tipo visual, lo cual tiene sentido ya que es en aquellas en las que la información visual y las decisiones tomadas desde el módulo CBIR más pueden ayudar a los resultados textuales, siempre y cuando se siga una estrategia de combinación adecuada.

7.5 Experimentación en corpus Buscamedia

Este apartado está dedicado a describir parte del trabajo experimental llevado a cabo dentro del proyecto Buscamedia, en relación a la recuperación de información multimedia y al tratamiento de consultas multimodales.

7.5.1 Fusión temprana de anotaciones multimedia

Se muestra en primer lugar un conjunto de experimentos destinados a evaluar el rendimiento de la fusión temprana de las anotaciones multimedia, así como la influencia de cada uno de los modos (vídeo, texto, audio, etc.) y fuentes de información (metadatos, transcripciones del audio, subtítulos en los vídeos, entidades nombradas identificadas, etc.) involucradas.

Para el desarrollo de estos experimentos se hace uso de la funcionalidad de *Búsqueda Configurable* (descrita en el apartado 4.3.5) implementada en Buscamedia. Esta funcionalidad

se encuentra disponible como servicio web en http://albali.lsi.uned.es/BMbusquedaConfigurableWS/, del que también existe una interfaz para su utilización en http://albali.lsi.uned.es/DemoBMbusquedaConfigurableWS/. La colección de evaluación utilizada es el corpus Deportes 20, desarrollado dentro del marco de trabajo del proyecto Buscamedia, y descrito en el apartado 4.2.2.

Como se describe en el apartado 4.3.5, la funcionalidad de *Búsqueda Configurable* permite seleccionar tanto los campos de búsqueda, como otros parámetros tales como el idioma, el operador lógico a utilizar o los tipos de recursos multimedia a recuperar. Ya que los experimentos están orientados a evaluar el rendimiento de la fusión temprana de anotaciones multimedia descrita en el apartado 4.2.2, estos son todos configurados de igual manera para el resto de parámetros: se utiliza el operador lógico OR, se recuperan todo tipo de recursos multimedia, y se establece como idioma por defecto el castellano al ser el mayoritario en la colección. Los parámetros configurables se corresponden con los campos del índice utilizados para llevar a cabo el proceso de búsqueda. Se analizará en primer lugar la aportación de cada uno de estos campos analizando el rendimiento del proceso de búsqueda cuando se utilizan todos los campos disponibles, y cuando se elimina cada uno de manera independiente. La siguiente tabla muestra los resultados obtenidos:

Tabla 7-44. Análisis fusión temprana de anotaciones multimedia (por fuentes)

metadatos	transcripciones	subtítulos	texto sobreimpreso	logos	objetos	entidades	MAP	Rprec	Recall
✓	-	-	-	-	-	-	0,4600	0,4326	70,18%
✓	✓	✓	✓	✓	✓	✓	**0,6936**	0,6436	89,57%
✓	✓	✓	✓	✓	✓	-	0,6620	0,6137	89,57%
✓	✓	✓	✓	✓	-	✓	0,6871	0,6357	89,57%
✓	✓	✓	✓	-	✓	✓	0,6740	0,6227	89,57%
✓	✓	✓	-	✓	✓	✓	0,6926	0,6436	89,57%
✓	✓	-	✓	✓	✓	✓	0,6745	0,6256	88,70%
✓	-	✓	✓	✓	✓	✓	**0,6898**	**0,6540**	89,57%

Para el análisis de rendimiento se utilizan las medidas de evaluación MAP (*Mean Average Precision*), Rprec (*Precision at R*, según el número R de elementos relevantes para cada consulta) y Recall, descritas en el apartado 4.1.1. No se utilizan la medidas de precisión a bajo nivel (*early precisions*), como P@5, P@10 y P@20, ya que la mayoría de las consultas no tienen suficientes documentos relevantes asociados (ver juicios de relevancia en apartado 4.2.2) para justificar el uso de este tipo de medidas. Por ejemplo, si para una consulta con solo 2 documentos relevantes en la colección se obtiene una lista de resultados ideal (los dos documentos relevantes en las dos primeras posiciones de la lista), la medida P@5 sería igual a 0,4 (2/5), $P@10 = 0{,}2$ y $P@20 = 0{,}1$, lo que no resultaría muy indicativo. En su lugar se utiliza la medida de evaluación Rprec, que calcula la precisión obtenida hasta los primeros R documentos recuperados para cada consulta, siendo R el número de documentos relevantes para dicha consulta. De esta manera se evita el problema planteado.

Puede observarse como los resultados obtenidos en base únicamente a los metadatos de los elementos del corpus obtienen unos valores de precisión y cobertura menores que los casos en los que se utilizan todas las anotaciones multimedia disponibles. Mientras que en el primer caso (solo metadatos) se obtiene un $MAP = 46\%$ y $Rprec = 43{,}26\%$, en el segundo (toda la información multimedia) se llega a un $MAP = 69{,}36\%$ y $Rprec = 64{,}36\%$. En cuanto a los valores de cobertura, estos son iguales o bastante parecidos en todos los casos en los que se utilizan la información multimedia ($MAP \approx 90\%$), siempre bastante mayor que cuando solo se emplean los metadatos ($MAP \approx 70\%$). Esto indica que las consultas de evaluación incluyen información referente a las características multimedia identificadas en los procesos de anotación.

El hecho de que los valores de evaluación obtenidos para los diferentes experimentos relacionados con cada uno de los tipos de información multimedia disponibles sean tan parecidos, puede ser debido a que las consultas fueron diseñadas para probar cada una de esas fuentes de información y, por lo tanto, que los documentos multimedia correctamente recuperados (verdaderos positivos) sean similares para cada caso. Entre las ligeras diferencias que se observan destaca el hecho de que la fuente de información cuya ausencia significa menos pérdida de precisión es la correspondiente a las transcripciones, con un $MAP = 0{,}6898$, lo cual tiene sentido debido a la poca información presente en dicha fuente (solo para 10 de los vídeos del corpus).

La fuente de información cuya presencia supone mayor ganancia en cuanto a precisión es la de las entidades nombradas. Cuando este campo no es utilizado para las búsquedas, los valores de precisión obtenidos son $MAP = 66{,}20\%$ y $Rprec = 61{,}37\%$, los menores de entre todos los experimentos realizados.

Con el objetivo de comparar la aportación de cada una de las modalidades de información multimedia disponibles en el corpus (texto, audio y vídeo), se agrupan las distintas fuentes de información según del modo del que provengan:

- TEXTO: metadatos, entidades nombradas
- AUDIO: transcripciones
- VÍDEO: subtítulos, texto sobreimpreso, logos, objetos

La siguiente tabla muestra los resultados obtenidos tras la evaluación según esta agrupación basada en la modalidad de la información utilizada:

Tabla 7-45. Análisis fusión temprana de anotaciones multimedia (por modo)

TEXTO	AUDIO	VÍDEO	MAP	Rprec	Recall
✓	✓	✓	**0,6936**	0,6436	89,57%
✓	✓	-	0,6185	0,5490	88,70%
✓	-	✓	**0,6898**	**0,6540**	89,57%
-	✓	✓	0,2272	0,2327	21,74%
✓	-	-	0,5884	0,5438	85,22%
-	✓	-	0,1590	0,1744	17,39%
-	-	✓	0,2352	0,2327	21,74%

Analizando los tres últimos experimentos, en los que se utilizan de manera única e independiente la información de cada uno de los modos disponibles (texto, audio y vídeo), queda bastante claro que el modo textual es el que más información relevante aporta y con el que individualmente se obtienen mejores resultados: se obtiene un $MAP = 58{,}84\%$ sensiblemente mayor que los obtenidos desde la modalidad de vídeo ($MAP = 23{,}52\%$) o con el audio ($MAP = 15{,}90\%$). El tipo de información que menos aporta es la correspondiente al audio, que incluso empeora en algunos casos los resultados de precisión cuando es utilizada ($Rprec = 64{,}36\%$ utilizando toda la información, y $Rprec = 65{,}40\%$ cuando se utiliza todo menos el audio). Esto puede ser debido a que la calidad de las transcripciones asociadas al modo audio no es de suficiente calidad, o que introduce demasiada información ruidosa, o como antes, que hay pocos documentos con transcripción.

Por último, se compara el funcionamiento de la búsqueda basada en todos los campos de información multimedia disponibles, con el del activo correspondiente a la *Búsqueda Automática* descrito en el apartado 4.3.5.

Tabla 7-46. Análisis fusión temprana de anotaciones multimedia (*Búsqueda Automática*)

metadatos	transcripciones	subtítulos	texto sobreimpreso	logos	objetos	entidades	MAP	Rprec	Recall
✓	✓	✓	✓	✓	✓	✓	**0,6936**	0,6436	89,57%
Búsqueda Automática							**0,7848**	**0,7301**	95,65%

Se ve cómo haciendo uso del activo de *Búsqueda Automática* y de sus reglas (descritas en el apartado 4.3.5) para seleccionar los campos de búsqueda más apropiados en cada caso, se obtienen mejores resultados que cuando se utilizan todos los campos de información disponibles, tanto para valores de precisión (MAP y Rprec) como de cobertura (Recall).

La siguiente gráfica muestra una comparativa entre los resultados de precisión (MAP) obtenidos haciendo uso de las distintas modalidades de información (Tabla 7-45), y aquellos correspondientes al caso de utilizar el activo de *Búsqueda Automática*.

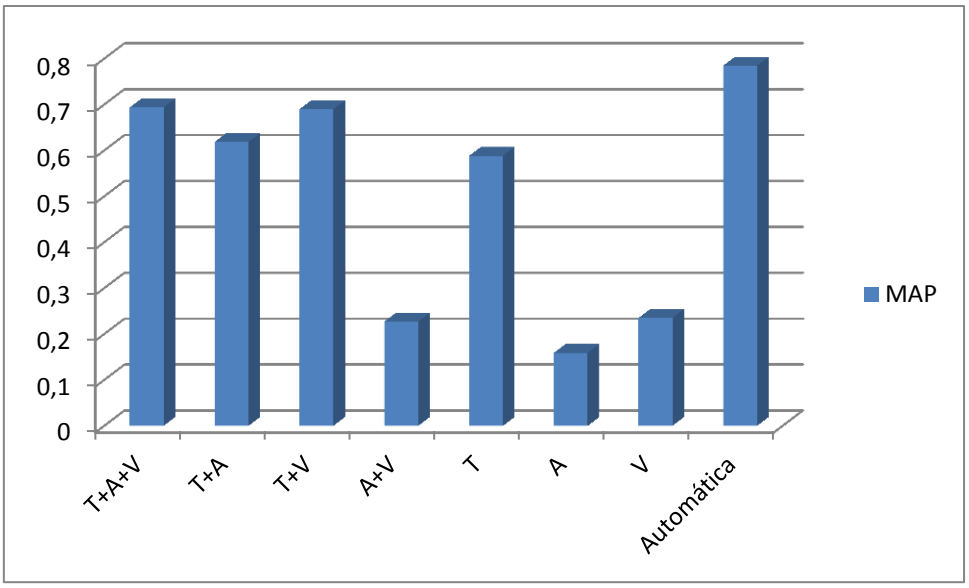

Figura 7.10. Fusión temprana de anotaciones multimedia (*Búsqueda Automática* VS por modalidades)

Los valores de precisión más altos se obtienen cuando la búsqueda es configurada por el activo de *Búsqueda Automática* en base a sus reglas. Esto muestra cómo, en algunos casos, la opción por defecto de utilizar toda la información disponible no resulta la más beneficiosa, ya que esta puede introducir ruido o simplemente reducir la precisión de los resultados.

7.5.2 Consultas multimodales

Dentro del proyecto Buscamedia se incluye la posibilidad de trabajar tanto con consultas textuales, como de incluir audio e imágenes. Para ello se proponen dos iniciativas: la primera basada en la forma de referenciar elementos multimedia en una consulta de texto, y la segunda relacionada con el tratamiento de las transcripciones de audio para mejorar el reconocimiento de nombres de entidades nombradas (personas, lugares u organizaciones) que puedan mencionarse en una consulta.

En cuanto al tratamiento de imágenes en consultas, se incluye una propuesta para el procesamiento de las mismas como parte de una consulta de usuario, prestando especial atención a los métodos de combinación de resultados en una fusión tardía. Esos resultados provendrán de llevar a cabo procesos de búsqueda tomando una imagen como punto de partida (y que devolverán imágenes como resultados), y procesos basados en texto (que devolverán documentos textuales o imágenes anotadas como respuesta).

7.5.2.1 Consultas multimodales formadas por texto e imágenes

Este apartado se centrará en las consultas multimedia o multimodales que involucren las modalidades de texto y de imagen (visual). Ejemplos de este tipo de consultas serían:

- TXT: solo texto. El tratamiento de la consulta textual se puede llevar a cabo aplicando un análisis de la consulta como el descrito en el apartado 4.3.5 en relación a la *Búsqueda Automática* implementada en Buscamedia.

- IMG: solo imagen. La consulta incluirá una imagen a partir de la cual se desean obtener otras imágenes relacionadas. Para ello podrán utilizarse tanto sistemas de recuperación visual (CBIR o *Query by Example*), descritos en el apartado 2.2.2, como técnicas TBIR basadas en la anotación de la imagen que forma parte de la consulta (por ejemplo detección de caras u objetos dentro de la imagen).

- TXT + IMG: texto e imagen. Cuando la consulta multimedia está formada por una parte visual (imagen/es) y otra textual, la estrategia de recuperación puede combinar la información disponible desde ambas modalidades. Para esto se utilizan técnicas de fusión a nivel de decisión o *late fusion* (descritas en el apartado 3.2.2), que combinan las listas de resultados obtenidas desde cada uno de los modos.

El esquema propuesto para el tratamiento de las consultas multimodales compuestas por texto e imágenes dentro del marco de trabajo del proyecto Buscamedia se muestra en la siguiente figura. En ella, la modalidad de audio también se incluye y es tratada, a partir de su transcripción, como información textual.

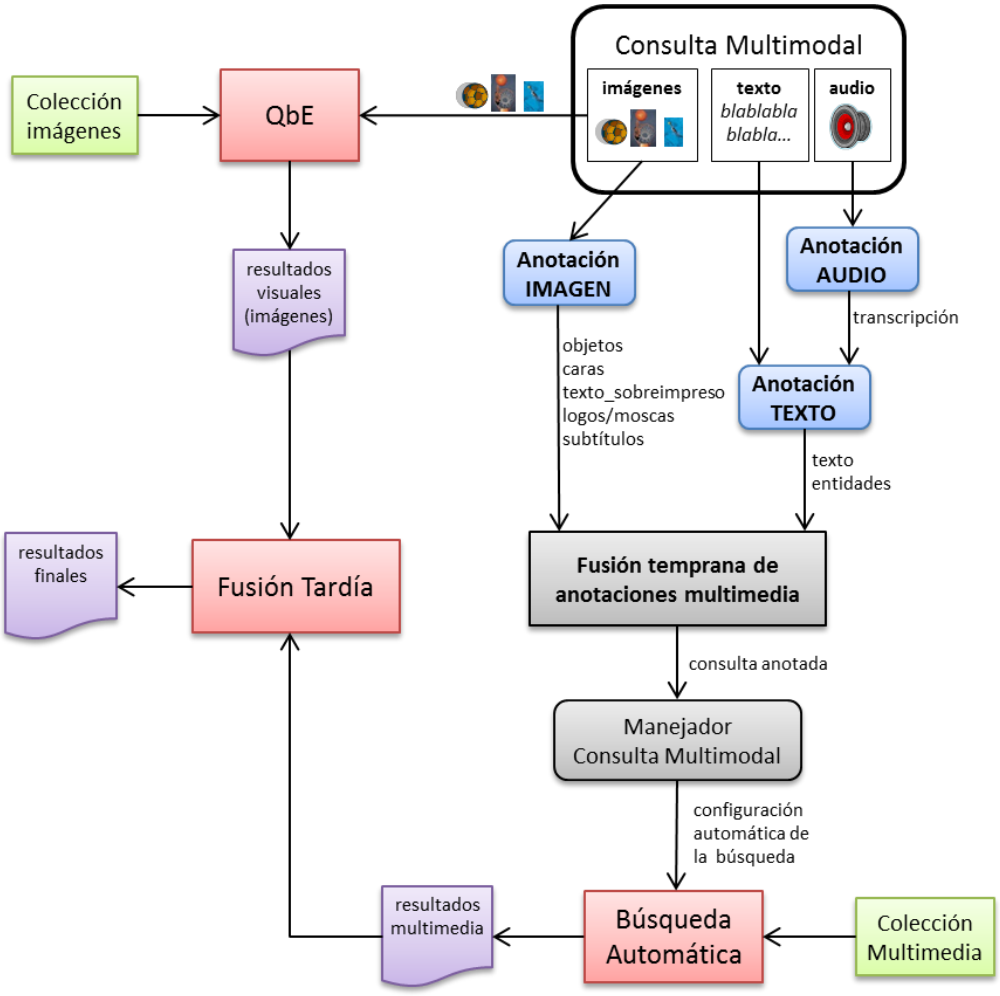

Figura 7.11. Tratamiento de consultas multimodales en Buscamedia

El proceso propuesto para tratar una consulta multimodal consiste básicamente en combinar mediante alguna aproximación de fusión tardía (*late fusion*) los resultados obtenidos mediante los sistemas de recuperación monomodal basados en: 1) texto, mediante la funcionalidad de *Búsqueda Automática*, y 2) imagen, mediante técnicas de *Query by Example* o CBIR (*Content-Based Image Retrieval*). Se describen a continuación cada una de estas alternativas:

1) Búsqueda textual automática. Se parte de la anotación textual de cada una de las modalidades de información presentes en la consulta multimedia:

   - imágenes → se anotan las imágenes de consulta haciendo uso de los distintos activos de anotación desarrollados en el proyecto, con los que se podrán identificar distintos elementos como: objetos, caras, logos, moscas, texto sobreimpreso o subtítulos.

   - texto → aparte del propio texto introducido por el usuario, se hace uso del activo de reconocimiento de entidades nombradas para identificar las entidades (de persona, organización, o lugar) presentes en el texto.

   - Audio → esta modalidad es utilizada como entrada para el activo de reconocimiento del habla (ASR, *Audio Speech Recognition*) para obtener la transcripción de audio proporcionado como parte de la consulta multimodal. La transcripción obtenida será tratada como texto normal y anotada como tal.

   Las distintas anotaciones procedentes de cada uno de los modos serán combinadas mediante una aproximación basada en fusión temprana o *early fusion* que será analizada por la funcionalidad de *Búsqueda Automática* (descrita en el apartado 4.3.5), la cual configurará los parámetros correspondientes antes de lanzar el proceso de búsqueda textual.

   La salida de este proceso de búsqueda basada en texto será una lista de resultados formada por recursos multimedia recuperados desde la colección.

2) Búsqueda visual (QbE). En este caso se parte de las imágenes proporcionadas como parte de la consulta multimodal. Estas imágenes de ejemplo son utilizadas por un

sistema de recuperación visual (*Query by Example* o CBIR) como los descritos en el apartado 2.2.2, que generará la correspondiente lista de imágenes recuperadas.

Una vez que los procesos de recuperación monomodales, basados en texto y en imagen respectivamente, han generado su lista de resultados, estos son combinados en una única lista final de recursos multimedia siguiendo una aproximación de fusión tardía o a nivel de decisiones (*late fusion*), como las descritas en el apartado 3.2.2.

Como parte del proyecto Buscamedia, se ofrece un servicio web que permite combinar listas de resultados siguiendo una aproximación basada en fusión tardía (*late fusion*). Dicho servicio trabaja a partir de las listas provenientes de diferentes orígenes y permite seleccionar el algoritmo de fusión tardía a aplicar. Actualmente el servicio permite seleccionar el formato de definición de las listas de resultados (TREC o Buscamedia), así como entre dos algoritmos de fusión distintos (*MaxMerge* y *Enrich*). El servicio se ofrece en la dirección http://albali.lsi.uned.es/BMfusionWS/, donde se describe su funcionalidad y parámetros de entrada, así como un ejemplo de llamada ().

BMfusion Web Service

Servicio Web que proporciona la funcionalidad de fusion.

Web Service

- http://albali.lsi.uned.es/BMfusionWS/BMfusionWS
- WSDL
- SOAP Request/Response Schema

Uso

Proporciona los siguientes métodos:

- *enrich (list1: String, list2: String, factor: float, norm: boolean, format: String): String*
  - Método que combina las listas de resultados principal (list1) y secundaria (list2) haciendo uso del algoritmo de fusion tardia Enrich.

Parámetros

- *list1:String* - Contenido de la lista de resultados principal.
- *list2:String* - Contenido de la lista de resultados secundaria.
- *factor:float* - Factor de enriquecimiento (Enrich).
- *norm:boolean* - Normalizar o no listas de resultados antes del proceso de fusion.
- *format:String* - Formato de las listas de resultados de entrada: "trec" o "buscamedia".

Se obtendrá como salida un String con el resultado de la fusion en formato TREC.

Figura 7.12. Servicio Web de fusión en Buscamedia

PARTE 3: CONCLUSIONES

Se incluyen en esta tercera parte las principales aportaciones obtenidas como fruto de la realización de esta tesis, así como las publicaciones relacionadas y las posibles líneas futuras de trabajo.

Capítulo 8 Principales aportaciones

La gran cantidad de información multimedia (imágenes, vídeos, etc.) existente en la actualidad en todo tipo de escenarios (colecciones personales, redes sociales, webs dedicadas a compartir documentos, etc.) hace necesario el avance en sistemas de almacenaje y búsqueda de este tipo de material multimedia. En este trabajo se investiga en la utilización de la fusión multimedia con el objetivo de aprovechar los distintos modos de información presentes en un escenario de recuperación de imágenes anotadas.

La aportación principal de esta tesis es la propuesta de una estrategia de fusión multimedia que explota las particularidades y la complementariedad entre las distintas modalidades de información de un objeto multimedia, con el objetivo de mejorar la calidad y la eficiencia de la tarea de recuperación. La técnica propuesta (LSMF, *Late Semantic Multimedia Fusion*) es evaluada dentro del escenario de la recuperación multimedia de imágenes, haciendo uso de distintos tipos de colecciones, principalmente aquellas proporcionadas dentro del foro de evaluación *ImageCLEF* (en el que se ha participado desde la edición de 2008).

La revisión del estado del arte, y la parte de la experimentación realizada en esta tesis, deja constancia del mejor rendimiento de las aproximaciones basadas únicamente en texto sobre

aquellas que solo utilizan información visual (descriptores o características de bajo nivel, como el color, la forma o la textura). La información de tipo textual es capaz de capturar el significado de las consultas y el de las imágenes, debido a la mayor carga semántica presente en dicho tipo de información en relación con la visual (*multimedia semantic gap*) por la complejidad para representar la información que un usuario percibe a partir de las características de bajo nivel de un objeto multimedia.

La investigación llevada a cabo confirma que una adecuada colaboración entre la información de tipo textual y la visual resulta beneficiosa en una tarea de recuperación multimedia. La complementariedad existente entre las distintas modalidades ayuda a reducir los efectos del problema del *semantic gap*. Diferentes algoritmos de fusión multimedia tardía (a nivel de decisiones) son implementados y evaluados, y los resultados obtenidos muestran como la combinación de las listas de resultados monomodales (TBIR y CBIR) genera listas fusionadas que mejoran la precisión de la recuperación. Por otro lado, la técnica propuesta simplifica el costoso proceso CBIR gracias a la reducción de la colección original en base a las anotaciones textuales de las imágenes, haciendo la tarea de recuperación escalable sobre grandes colecciones multimedia. Esta reducción es llevada a cabo mediante la aplicación de un prefiltro basado en la información textual asociada a las imágenes. Este prefiltro restringe el conjunto de imágenes a un subconjunto formado únicamente por aquellas que guardan algún tipo de relación semántica/textual con las consultas.

Como parte de la investigación relacionada con la recuperación de imágenes basada únicamente en las anotaciones textuales asociadas a las imágenes (TBIR), se llevan a cabo un conjunto de experimentos para obtener una óptima configuración del sistema. Se evalúan aspectos relacionados con el preprocesamiento textual y con la recuperación multilingüe, comparando técnicas de fusión textual (tanto a nivel de características, como a nivel de decisiones). Los resultados muestran cómo la aproximación que mejor funciona para el multilingüismo es la basada en una fusión monomodal a nivel de decisiones (*late fusion*), esto es, generar listas de resultados independientes para cada uno de los idiomas disponibles en la colección para, posteriormente, combinarlas en una única lista final fusionada multilingüe. En cuanto al preprocesamiento, también se recomienda la aplicación de técnicas de *stemming* y eliminación de *stopwords*. El análisis relacionado con el reconocimiento de entidades nombradas no mejora significativamente los resultados de la recuperación, debido

principalmente a la poca presencia de estas en las colecciones y las consultas de evaluación utilizadas.

Siguiendo con la recuperación textual, se comprueba que el enriquecimiento en base a Wikipedia como recurso externo mejora el rendimiento del sistema, incrementando la precisión de los resultados. El enriquecimiento consiste en ampliar las anotaciones de las imágenes de la colección extrayendo información textual adicional de los artículos de Wikipedia en los que aparecen las imágenes.

En lo que se refiere a la recuperación basada en las características visuales (CBIR), se analiza la posibilidad de llevarla a cabo en base a grupos independientes de descriptores visuales (color, textura, etc.), en comparación con utilizar todos ellos de manera conjunta. La conclusión a la que se llega, en base a los resultados obtenidos, es que el modelo de realimentación por relevancia basado en regresión logística utilizado para la recuperación visual, funciona mejor cuando se utilizan todos los descriptores visuales de forma conjunta, y no separándolos en grupos independientes. Tampoco la combinación multimedia entre estos resultados CBIR (por grupos de descriptores) y los obtenidos desde TBIR, siguiendo la estrategia LSMF, consigue mejorar a aquellos obtenidos cuando se trabaja desde el principio con todos los descriptores visuales conjuntamente.

Como parte del análisis de la técnica de LSMF, se evalúa inicialmente la fase de prefiltrado textual. Los experimentos llevados a cabo muestran el buen funcionamiento del mismo, logrando una gran reducción de la base de datos original (casi un 98%), sin reducir demasiado la cobertura del subconjunto de imágenes generado (casi un 83%). La conclusión obtenida es que las características textuales captan adecuadamente el significado semántico de las consultas de los usuarios, lo que puede ser utilizado para simplificar la colección de imágenes sobre la que trabajará el sistema CBIR. De este modo los resultados visuales son ampliamente mejorados, a la vez que se reduce el tiempo de proceso de CBIR, lo que permite hacer la tarea de recuperación escalable sobre grandes colecciones multimedia.

En cuanto a la comparación entre los esquemas clásicos de fusión multimedia a nivel de decisiones (LF) y la técnica LSMF propuesta en esta tesis, se comprueba que los resultados obtenidos en base a la segunda aproximación son mejores tanto en términos de precisión global como en la de las imágenes recuperadas en las primeras posiciones (*early precision*).

Esto confirma la influencia positiva del prefiltrado textual realizado previamente a la recuperación visual, que permite al sistema CBIR trabajar sobre un conjunto reducido de imágenes que guardan relación semántica/textual con las consultas, y disponer de los contraejemplos visuales (ejemplos de imágenes negativos para la consulta) necesarios para el algoritmos de recuperación visual de realimentación por relevancia basado en regresión logística. Además de la mejora de la calidad de la recuperación, la simplificación de la colección original reduce significativamente el tiempo de proceso CBIR, y mejora sus resultados. La combinación entre los resultados visuales (CBIR) y los textuales (TBIR) aumentará la precisión de los resultados.

Los algoritmos de fusión tardía que mejor rendimiento ofrecen, tanto dentro de la técnica LSMF propuesta como en un esquema cásico (LF), son el *Product* y el *OWA*, con mejoras cercanas al 18% sobre TBIR en alguno de los casos. El único algoritmo que utiliza para la combinación multimedia solo la información referente al *rank* (posición ocupada por las imágenes en las listas de resultados) es el *FilterN*, que obtiene resultados muy bajos cuando se trabaja sobre la colección completa (sin prefiltrar), debido a que se eliminan demasiadas imágenes relevantes por la poca precisión de los resultados visuales. Cuando *FilterN* trabaja sobre la colección prefiltrada (LSMF) los resultados son muy similares a los de TBIR, incluso ligeramente mejores en P@5 para valores restrictivos de N. En cuanto al algoritmo *Enrich*, que utiliza tanto el valor de relevancia o *score* como el *rank* obtenido por las imágenes en las listas monomodales, la diferencia de rangos de los *scores* hace que el peso de la lista CBIR no sea muy influyente en los resultados finales. Por este motivo, el rendimiento es muy similar al de TBIR, aunque se obtienen ligeras mejoras.

Las comparaciones llevadas a cabo entre los resultados obtenidos por la propuesta de LSMF y otras aproximaciones observadas en el foro de evaluación *ImageCLEF*, confirman el buen comportamiento de la técnica propuesta en esta tesis. Durante las dos últimas ediciones (años 2010 y 2011) de la tarea de recuperación multimedia de imágenes de Wikipedia, LSMF obtenidos un rendimiento superior al de todos los demás grupos participantes. Únicamente el grupo de Xerox obtiene unos resultados similares, en base a una aproximación similar. Por lo tanto, puede concluirse que una estrategia de combinación multimedia basada en una fase inicial de prefiltrado textual (capaz de restringir la colección original de búsqueda a imágenes

que guarden alguna relación semántica con las consultas), seguida de la fusión de los resultados monomodales, resulta adecuada en una tarea de recuperación multimedia.

Los experimentos llevados a cabo para analizar la conveniencia de incluir una fase de normalización de los *scores* de las listas de resultados monomodales antes del proceso de fusión, muestran que esta no es necesaria. Es más, para todos los algoritmos de fusión evaluados, y para casi todas sus configuraciones, los resultados empeoran cuando se normalizan los *scores*. Solo para algunos casos concretos con el algoritmo *OWA,* los resultados mejoran ligeramente. De todo esto, se puede extraer que no es necesario incluir la fase de normalización para adaptar todos los *scores* dentro de un mismo rango de valores, ya que la aproximación basada en confiar en los valores generados desde cada subsistema de recuperación obtiene mejores resultados de fusión.

Los análisis realizados acerca de la fase de prefiltrado textual y de la estrategia LSMF completa son llevados a cabo adicionalmente en función de la complejidad y de la carga visual de las consultas multimedia. Los resultados obtenidos muestran cómo los valores de cobertura y precisión alcanzados son mayores cuanto más sencillas son las consultas (casi un 90% de cobertura con las consultas clasificadas como fáciles). En cuanto a la carga visual, son las consultas de tipo neutro las que mejores valores de cobertura y precisión alcanzan. Por otro lado, la técnica LSMF consigue las mejoras (en relación a TBIR) más significativas para las consultas más difíciles, principalmente en los valores de precisión a bajo nivel. Esto es debido a que las consultas más fáciles ya obtienen buenas respuestas en base a la recuperación textual monomodal. Esto se corrobora con los resultados de mejora observados en función de la carga visual de las consultas, que muestra incrementos para los tipos visual y neutral, pero no el textual, ya que para dichas consultas la parte visual de las mismas no aporta información relevante o complementaria a la textual. Por el mismo motivo, son las consultas de tipo visual las que mayor porcentaje de mejora experimentan cuando se aplica LSMF en relación a la aproximación TBIR.

Fuera de la tarea concreta de la recuperación de imágenes, dentro del proyecto Buscamedia, se ha construido una colección de objetos multimedia formada por vídeos, imágenes o *keyframes*, noticias, y páginas webs. Además, se han definido un conjunto de consultas, con sus correspondientes juicios de relevancia (o *ground truth*), con el que se evalúan aspectos relacionados con la fusión temprana de anotaciones multimedia. Los resultados obtenidos

confirman que la fusión temprana de anotaciones de los objetos multimedia, en base a otras fuentes de información (transcripciones de audio, texto sobreimpreso en vídeos o *keyframes*, subtítulos, etc.), resulta beneficioso para mejorar tanto la cobertura como la precisión de la recuperación multimedia. Se comprueba igualmente que la configuración automática de los parámetros y campos de búsqueda en base a un conjunto de reglas, funciona mejor que la selección manual de dichos parámetros por parte de los usuarios.

Los resultados obtenidos dentro de los escenarios de evaluación planteados, muestran el buen funcionamiento de la estrategia de fusión multimedia propuesta, tanto en términos de precisión de resultados como en relación al tiempo de búsqueda. Esta propuesta, aunque evaluada sobre un conjunto de imágenes anotadas, podría ser aplicada a cualquier otro tipo de colección multimedia en la que esté presente la modalidad textual.

Capítulo 9 Líneas futuras de trabajo

Durante el desarrollo de este trabajo de investigación se han realizado numerosos experimentos destinados a descartar aproximaciones o ideas poco prometedoras y a mejorar las técnicas propuestas. Sin embargo, hay algunas líneas de trabajo futuro que potencialmente pueden seguir mejorando el problema abordado en esta tesis y que resuelve, en cierta medida, la estrategia de fusión multimedia semántica tardía (LSMF) propuesta.

En primer lugar, la etapa de prefiltrado textual de la estrategia LSMF podría ser parametrizada. Los experimentos realizados consideran suficiente un valor de relevancia textual o *score* (S_t) mayor que cero, para que dicha imagen supere el prefiltro. ¿Qué ocurriría si el prefiltro es más restrictivo?, ¿mejoraría la precisión? ¿Y la relación entre cobertura y precisión? ¿Existe el riesgo de dejar fuera de la subcolección demasiadas imágenes relevantes?

Un aspecto muy interesante es explorar técnicas de expansión de la consulta, para aumentar la cobertura de la recuperación textual (TBIR). La expansión de la consulta podría llevarse a cabo a partir de la información textual, haciendo uso de ontologías de conceptos visuales o de recursos como WordNet. Pero además de identificar las imágenes de ejemplo y contraejemplo para la consulta multimodal como en la propuesta presentada en esta memoria, una expansión dirigida por características visuales, ¿hasta dónde permitiría mejorar? El análisis de la subcolección prefiltrada y de los resultados visuales (CBIR) obtenidos a partir de ella, ¿daría pistas sobre la conveniencia o no de hacer una expansión?

El enriquecimiento de la información textual planteado en esta tesis (haciendo uso de los artículos de Wikipedia en los que aparecen las imágenes), y cuya influencia positiva ha sido comprobada, podría ser refinado tratando de explotar otra información disponible en internet. Ante la ausencia o escasez de anotaciones textuales, ¿podría la información textual relacionada con párrafos cercanos (y otros textos) mejorar los resultados de búsqueda? Además, trabajando con características visuales ¿podrían detectarse en alguna medida los conceptos visuales presentes en las imágenes?

Aunque la estrategia de fusión multimedia (LSMF) propuesta en esta tesis es evaluada dentro de un escenario de recuperación de imágenes, resultaría interesante probar su funcionamiento sobre una colección multimedia de vídeos y textos de internet, (para ello sería necesario disponer de un conjunto de consultas y sus juicios de relevancia), colección más completa que con la que se ha intentado probar (la colección desarrollada en el proyecto de investigación Buscamedia). De este modo, se probarían las conclusiones de esta tesis en otros escenarios reales.

Capítulo 10 Producción científica

Se enumeran este capítulo las distintas publicaciones realizadas por el autor de esta tesis durante el desarrollo de la misma.

- Xaro Benavent, Ana Garcia-Serrano, **Ruben Granados**, Joan Benavent and Esther de Ves. *Multimedia Information Retrieval based on Late Semantic Fusion Approaches: Experiments on a Wikipedia Image Collection*. To appear in IEEE Transactions on Multimedia Journal. 2013. Impact Factor (2012): 1.754 (Q1: 15/105). DOI: 10.1109/TMM.2013.2267726. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6529192

- A. García-Serrano, **R. Granados**, D. Hernández-Aranda, V. Fresno, J. Cigarrán. *Anotación para la recuperación de información multimedia: el corpus Deportes20*. Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2012). Septiembre de 2012. http://krono.act.uji.es/SEPLN/p12.pdf.

- David Hernández-Aranda, **Rubén Granados**, A. García Serrano. *Servicios de anotación y búsqueda para corpus multimedia*. Procesamiento del Lenguaje Natural, nº49, 2012. (Revista con el sello de calidad ISO9001, y certificado de "Revista Excelente" otorgado por la FECYT en la III Convocatoria de Evaluación de la Calidad Editorial y Científica 2012). http://sinai.ujaen.es/sepln/ojs/ojs/index.php/pln/article/view/4568/2741.

- **Rubén Granados Muñoz**, Ana García-Serrano, Noelia Méndez Fernández, Xaro Benavent García. *Experimentación en la Búsqueda de Imágenes a partir de características Visuales y Textuales: Fusión Tardía y Expansión de la Consulta*. Procesamiento del Lenguaje Natural, nº 48, 2012, pp: 73-80. (Revista con el sello de calidad ISO9001, y certificado de "Revista Excelente" otorgado por la FECYT en la III Convocatoria de Evaluación de la Calidad Editorial y Científica 2012). ISSN (edición impresa): 1135-5948. ISSN (edición electrónica): ISSN: 1989-7553. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/download/4490/2658.

- **Rubén Granados**, Joan Benavent, Xaro Benavent, Esther de Ves, Ana García-Serrano: Multimodal Information Approaches for the Wikipedia Collection at ImageCLEF 2011. CLEF (Notebook Papers/Labs/Workshop). http://clef2011.org/resources/proceedings/Granados_Clef2011.pdf. In CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, the Netherlands. ISBN: 978-88-904810-1-7. 2011.

- Juan Manuel Cigarrán Recuero, Víctor Fresno Fernández, Ana M. García-Serrano, David Hernández Aranda, **Rubén Granados**. UNED at MediaEval 2011: Can Delicious help us to improve automatic video tagging? http://ceur-ws.org/Vol-807/Cigarran_UNED2011_Genre_me11wn.pdf. Working Notes Proceedings of the MediaEval 2011 Workshop, Santa Croce in Fossabanda, Pisa, Italy, September 1-2, 2011. CEUR Workshop Proceedings, Volume 807. 2011.

- David Hernández Aranda, Víctor Fresno Fernández, **Rubén Granados Muñoz**, Ana García Serrano. Evaluando Modelos de Indexación Multilingüe en Recuperación de Imágenes Basada en Texto. CAEPIA 2011. Tenerife, España. http://aepia.aic.uniovi.es/revista/index.php/aia/article/view/917/740.

- Rodrigo Agerri, **Rubén Granados**, Ana García Serrano. ***Enrichment of Named Entities for Image Photo Retrieval***. 7th International conference on Adaptive Multimedia Retrieval: Understanding Media and Adapting to the User. Lecture Notes

in Computer Science. Volume 6535. ISBN: 978-3-642-18448-2. pp: 101-110. Springer. 2011. http://www.springerlink.com/content/b871066l507q2j4x/.

- J. Benavent, X. Benavent, E. de Ves, **R. Granados**, A. García-Serrano. ***Experiences at ImageCLEF 2010 using CBIR and TBIR Mixing Information Approaches.*** CLEF (Notebook Papers/LABs/Workshops). http://clef2010.org/resources/proceedings/clef2010labs_submission_74.pdf. In CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy. ISBN: 978-88-904810-0-0. 2010.

- D. Hernández-Aranda, **R. Granados**, J. Cigarran, A. Rodrigo, V. Fresno, and A. García-Serrano. ***Uned at mediaeval 2010: exploiting text metadata for automatic video tagging***. In MediaEval 2010 Workshop. Pisa, Italy, 24 October, 2010. http://www.multimediaeval.org/worknotes2010/UNED_TaggingProf.pdf.

- Ana García-Serrano, Xaro Benavent, **Rubén Granados**, Esther de Ves, Jose Miguel Goñi. ***Multimedia Retrieval by Means of Merge of Results from Textual and Content Based Retrieval Subsystems***. Multilingual Information Access Evaluation II - Multimedia Experiments. Revised selected papers from 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009. Lecture Notes in Computer Science. Volume 6242/2010. DOI: 10.1007/978-3-642-15751-6. ISBN: 978-3-642-15750-9. pp: 142-149. #pp: 8. 2009. http://www.springerlink.com/content/542r181u253122p7/.

- **R Granados**, Ana García Serrano. ***Combinación de técnicas textuales y visuales para la recuperación de imágenes***. Actas de las III Jornadas sobre modelos y técnicas para el acceso a la información multilingüe y multimodal en la Web, 5 y 6 de febrero de 2009, UC3M, Colmenarejo, Madrid. http://sinai.ujaen.es/timm/jornadas_pln/presentaciones-timm09/Granados-RImagenes.ppt.pp: 75-76. #pp: 2. 2009.

- **Rubén Granados Muñoz**, Ana García Serrano, José M. Goñi Menoyo. ***La herramienta IDRA (Indexing and Retrieving Automatically)***. Procesamiento del Lenguaje Natural, nº 43, 2009. http://www.sepln.org/revistaSEPLN/revista/43/articulos/art40.pdf. XXV Conferencia de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN'09). San Sebastián, 2009. #pp: 2. 2009.

- Ana García-Serrano, Xaro Benavent, **Rubén Granados**, José Miguel Goñi-Menoyo. ***Some Results Using Different Approaches to Merge Visual and Text-Based Features in CLEF'08 Photo Collection***. Evaluating Systems for Multilingual and Multimodal Information Access. Revised selected papers from 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008. ISSN: 0302-9743 (Print) 1611-3349 (Online). Lecture Notes in Computer Science. Volume 5706/2009. ISBN: 978-3-642-04446-5. pp: 568-571. #pp: 4. 2009. http://www.springerlink.com/content/h8152q582351606t/.

Anexo. Herramienta IDRA

Se describe en este Anexo la primera versión de la herramienta IDRA (*InDexing & Retrieving Automatically*) (Granados Muñoz, García-Serrano and Goñi Menoyo, 2009) y sus principales funcionalidades.

INTRODUCCIÓN

La herramienta IDRA se desarrolla como parte del proceso de investigación de esta tesis con el objetivo de proporcionar las funcionalidades necesarias para el desarrollo de las diferentes experimentaciones llevadas a cabo. Se encuentra disponible como software libre en *SourceForge*[30] con el fin de que pueda ser utilizada por todos aquellos que la puedan necesitar. Está desarrollada completamente en Java, por lo que podrá ser descargada y utilizada en cualquier equipo informático. Hasta la redacción de esta tesis se han registrado más de 1.000 descargas.

Inicialmente IDRA nace para ofrecer a sus usuarios la posibilidad de indexar y buscar sobre todo tipo de documentos textuales (txt, doc, pdf, etc.). Poco a poco la herramienta fue creciendo y se incluyeron otras funcionalidades como la indexación de la parte textual de distintas colecciones multimedia de imágenes anotadas, la combinación o fusión de listas de resultados, o la representación de los contenidos indexados en cada momento, con el fin de facilitar el análisis de los mismos.

[30] http://sourceforge.net/

Figura 10.1. Herramienta IDRA

Se describen en los siguientes apartados la estructura interna de la información manejada por la herramienta IDRA, así como las principales funcionalidades proporcionadas (preprocesamiento textual, indexación, recuperación, gestión del contenido, evaluación, fusión, etc.).

ESTRUCTURAS DE DATOS

El modelo de indexación y recuperación de la herramienta IDRA está basado en el VSM (modelo del espacio vectorial, descrito en el apartado 2.1.1.2). Por ese motivo, es necesario almacenar cierta información acerca de cada documento y de su contenido textual. Para ello se implementa una estructura de datos propia desarrollada enteramente en Java.

Las estructuras de datos utilizadas manejan un conjunto de vectores que almacenan diferentes valores para cada uno de los términos de los documentos que se indexan. Cada uno de estos vectores contendrá la siguiente información:

- *palabras*. Cada uno de los términos representantes de los documentos indexados.
- *frecs*. Número de documentos en que aparece el término.
- *ids*. Identificador del documento al que pertenece el término.
- *apars*. Número de veces que el término aparece en el documento indexado.

- *idfs*. (Inverse document frequency) Mide la importancia general del término en la colección.

- *DE*. Distancia euclídea entre los pesos de los términos de un documento.

- *weights*. Peso de cada uno de los términos en su correspondiente documento.

- *similarity*. Valor de similitud entre un documento y la consulta introducida en ese momento.

Una vez indexada una determinada colección, estos valores no cambiarán (sólo el de '*similarity'*, que depende de las consultas introducidas), y se dispondrá de ellos en cualquier momento para realizar consultas y recuperar documentos sobre la colección indexada. Estas estructuras de datos pueden ser salvadas en cualquier momento, haciendo uso del módulo de gestión del contenido, y ser recuperadas posteriormente. También será posible generar tablas, en ficheros de texto o en bases de datos, para facilitar el análisis de la información.

FUNCIONALIDADES

IDRA está organizada en diferentes módulos, cada uno de ellos encargado de un determinado tipo de funcionalidades. Esta estructura puede verse siempre en la parte superior de la interfaz, desde donde se puede navegar por estos módulos a través de pestañas:

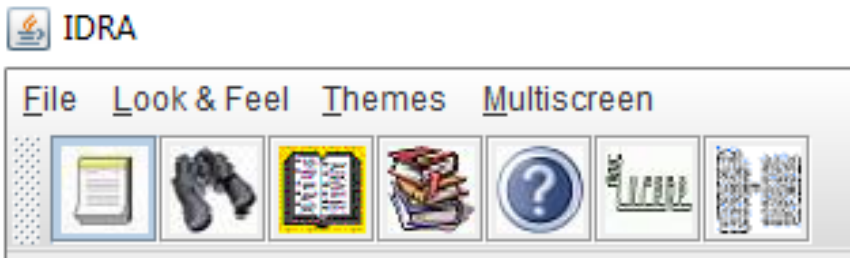

Figura 10.2. Funcionalidades principales de IDRA

Como puede observarse en la figura anterior, IDRA ofrece 7 módulos distintos cuya contenido se muestra en las siguientes secciones.

- **Módulo de indexación de documentos**

Permite al usuario la extracción del texto de documentos en diferentes formatos, el preprocesamiento de dicha información textual, y la indexación de documentos individuales o de colecciones completas.

La aproximación seguida para la representación, indexación y recuperación de los documentos está basada en el modelo del espacio vectorial, utilizando la función de pesado TF-IDF para calcular el valor de cada término en cada documento.

Figura 10.3. Módulo de indexación de documentos

- **Módulo de recuperación de documentos**

Permite realizar el mismo preprocesamiento llevado a cabo para los textos a indexar, pero ahora sobre el texto de las consultas. También se encarga de calcular la relevancia de los documentos indexados en relación a las consultas introducidas y, de este modo, presentar al usuario el conjunto de documentos recuperados ordenados.

Al igual que la indexación, el enfoque seguido es el basado en el VSM, y la función de similitud para calcular los valores de relevancia es el coseno.

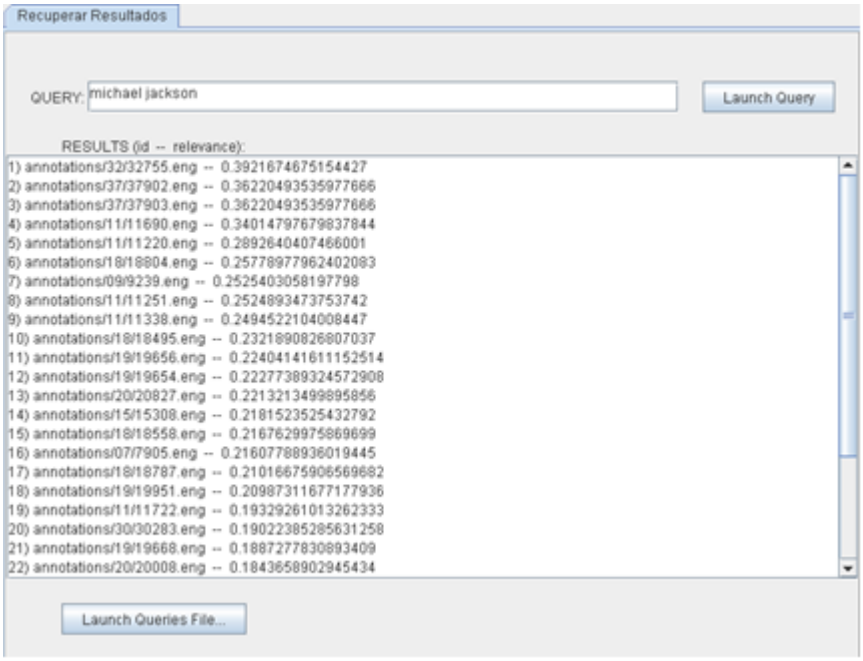

Figura 10.4. Módulo de recuperación de documentos

- **Módulo de Preparación de Datos**

Pensado para la preparación de distintos tipos de datos (colecciones de documentos, lista de consultas de evaluación, juicios de relevancia, etc.) para su posterior tratamiento por parte de la herramienta.

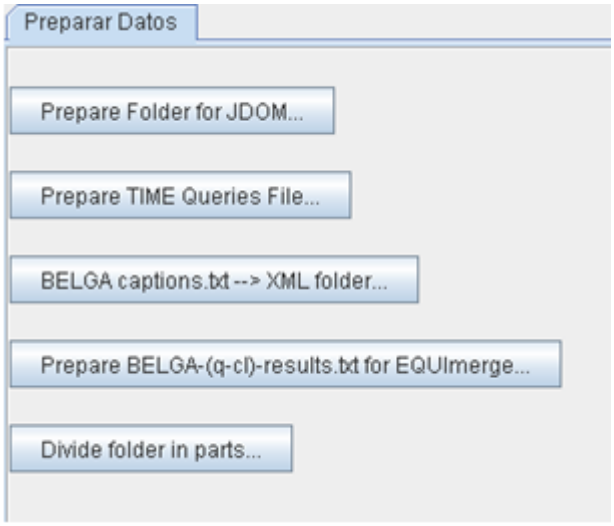

Figura 10.5. Módulo de Preparación de Datos

- **Módulo de gestión del contenido**

El usuario podrá manejar el índice a utilizar en cada momento: podrá inicializar, salvar o cargar el índice que desee. También ofrece la posibilidad de generar una tabla mostrando las características principales de cada indexación y de su representación interna (términos indexados, frecuencia de cada término, peso de cada término en cada documento, etc.).

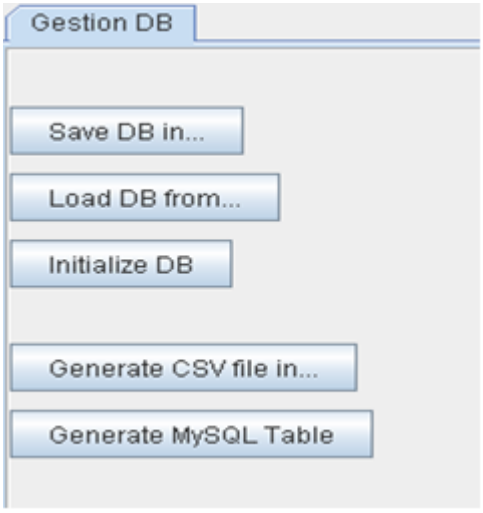

Figura 10.6. Módulo de gestión del contenido

- **Módulo de Evaluación de Resultados**

Se podrán calcular los valores de algunas de las medidas de evaluación más significativas dentro del campo de la recuperación de información (en formato TREC, o en el propio de IDRA). Esto facilitará el análisis de los resultados obtenidos por la herramienta para una determinada configuración o experimento.

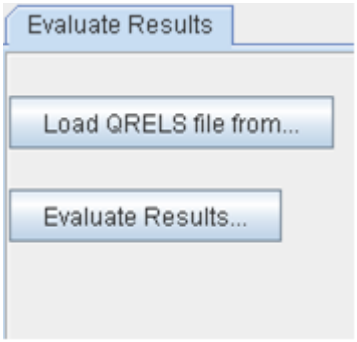

Figura 10.7. Módulo de Evaluación de Resultados

- **Módulo Lucene**

Haciendo uso de este módulo, el usuario podrá conectarse con las funcionalidades de indexación y búsqueda proporcionadas por la herramienta Lucene.

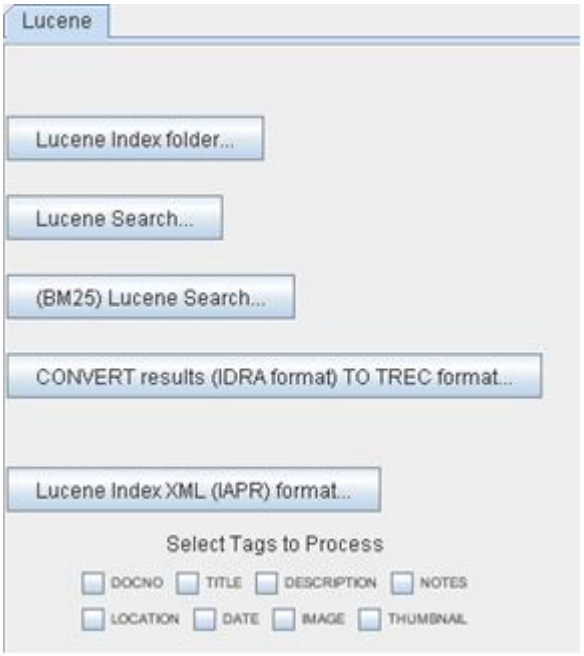

Figura 10.8. Módulo Lucene

- **Módulo de algoritmos de fusión**

Aquí se ofrece la implementación de varios algoritmos de fusión tardía (late fusion) o a nivel de resultados, que permitirán combinar varios resultados de recuperación y generar una única lista fusionada. Se encuentran implementados los algoritmos utilizados durante el desarrollo de la tesis: FilterN, Enrich, EquiMerge, MaxMerge, Product.

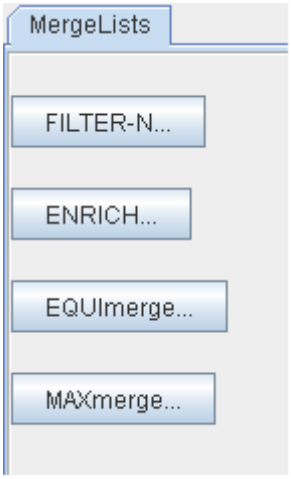

Figura 10.9. Módulo de algoritmos de fusión

Bibliografía

Adams, W., Iyengar, G., Lin, C., Naphade, M., Neti, C., Nock, H. and Smith, J. (2003) 'Semantic indexing of multimedia content using visual, audio, and text cues', *EURASIP J. Appl. Signal Process*, pp. 170-185.

Agerri, R., Granados, R. and Garcia-Serrano, A. (2011) 'Enrichment of Named Entities for Image Photo Retrieval', in *Adaptive Multimedia Retrieval. Understanding Media and Adapting to the User*.

Ah-Pine, J., Bressan, M., Clinchant, S., Csurka, G., Hoppenot, Y. and Renders, J.-M. (2009) 'Crossing textual and visual content in different application scenarios', *Multimedia Tools and Applications*, vol. 42, pp. 31-56, Available: 10.1007/s11042-008-0246-8, 1380-7501.

Arampatzis, A., Chatzichristofis, S.A. and Zagoris, K. (2010) 'Multimedia Search with Noisy Modalities: Fusion and Multistage Retrieval', in *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy.

Arampatzis, A., Konstantinos, Z. and Chatzichristofis, S.A. (2011) 'DUTH at ImageCLEF 2011 Wikipedia Retrieval', in *CLEF 2011 working notes*.

Arni, T., Clough, P., Sanderson, M. and Grubinger, M. (2009) 'Overview of the ImageCLEFphoto 2008 Photographic Retrieval Task', in *Evaluating Systems for Multilingual and Multimodal Information Access*, Springer.

Aslam, J.A. and Montague, M. (2001) 'Models for metasearch', SIGIR'01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 276-284.

Atrey, P., Hossain, M. and Kankanhalli, M. (2010) 'Multimodal fusion for multimedia analysis: a survey', *Multimedia Systems*, pp. 345-379.

Awadi, H., Khemakhem, M.T. and Jemaa, M.B. (2011) 'Evaluating some contextual factors for image retrieval: ReDCAD participation at ImageCLEFWikipedia 2011', in *CLEF 2011 working notes*.

Baeza-Yates, R.A. and Ribeiro-Neto, B. (1999) *Modern Information Retrieval*, Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Baeza-Yates, R. and Ribeiro-Neto, B. (2011) *Modern Information Retrieval (second edition)*, Addison-Wesley.

Bell, A.J. (2003) 'The co–information lattice', Proceedings of the 4th international symposium on independent component analysis and blind signal separation (ICA2003), Nara, Japan, 921–926.

Benavent, J., Benavent, X., de Ves, E., Granados, R. and Garcia-Serrano, A. (2010) 'Experiences at ImageCLEF 2010 using CBIR and TBIR Mixing Information Approaches', in Braschler, M., Harman, D. and Pianta, E. *CLEF 2010 LABs and Workshops, Notebook Papers*, Padoua, Italy.

Benavent, J., Benavent, X., Granados, R. and Garcia-Serrano, A. (2010) 'Experiences at ImageCLEF 2010 using CBIR and TBIR Mixing Information Approaches', in *CLEF 2010 Working Notes*.

Benavent, X., Garcia-Serrano, A., Granados, R., Benavent, J. and de Ves, E. (2013) 'Multimedia Information Retrieval based on Late Semantic Fusion Approaches: Experiments on a Wikipedia Image Collection', *IEEE Transactions on Multimedia*.

Berber, T. and Alpkocak, A. (2009) 'DEU at ImageCLEFmed 2009: Evaluating re–ranking and integrated retrieval model', in *Working Notes of CLEF 2009*, Corfu, Greece.

Berber, T., Vahid, A.H., Ozturkmenoglu, O., Gachpaz, R. and Alpkocak, A. (2011) 'DEMIR at ImageCLEFwiki 2011: Evaluating Different Weighting Schemes in Information Retrieval', in *CLEF 2011 working notes*.

Boros, E., Ginsca, A.L. and Iftene, A. (2011) 'UAIC's participation at Wikipedia Retrieval @ ImageCLEF 2011', in *CLEF 2011 working notes*.

Buckley, C. and Voorhees, E. (2000) 'Evaluating evaluation measure stability', Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, Athens, Greece, 33--40.

Caicedo, J.C., Moreno, J.G., Niño, E.A. and González, F.A. (2010) 'Combining visual features and text data for medical image retrieval using latent semantic kernels', Proceedings of the international conference on Multimedia information retrieval, Philadelphia, Pennsylvania, USA, 359--366.

Cao, L., Chang, Y.C., Codella, N., Merler, M., Nguyen, Q.B. and Smith, J.R. (2012) 'IBM T.J. Watson Research Center, Multimedia Analytics: Modality Classification and Case-based Retrieval tasks of ImageCLEF 2012', ImageCLEF 2012 Working Notes.

Carlsson, C. and Fuller, R. (2002) *Fuzzy Reasoning in Decision Making and Optimization*, Springfield-Verlag.

Carpineto, C. and Romano, G. (2012) 'A Survey of Automatic Query Expansion in Information Retrieval', *ACM Comput. Surv.*, vol. 44, no. 1, pp. 1-50, Available: 0360-0300, 10.1145/2071389.2071390.

Chaisorn, L., Chua, T.S., Lee, C.H., Zhao, Y., Xu, H., Feng, H. and Tian, Q. (2003) 'A multi-modal approach to story segmentation for news video', *World Wide Web*, pp. 187-208.

Chatzichristofis, S.A., Zagoris, K., Boutalis, Y.S. and Papamarkos, N. (2010) 'Accurate Image Retrieval based on Compact Composite Descriptors and Relevance Feedback Information', *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, vol. 24, no. 2, February, pp. 207-244.

Chowdhury, G. (2010) *Introduction to Modern Information Retrieval, Third Edition*, Facet Publishing.

Cigarrán Recuero, J.M., Fernandez, V.F., Garcia-Serrano, A., Aranda, D.H. and Granados, R. (2011) 'UNED at MediaEval 2011: Can Delicious help us to improve automatic video tagging?', Working Notes Proceedings of the MediaEval 2011 Workshop.

Cleverdon, C.W. (1960) 'The ASLIB Cranfield research project on the comparative efficiency of indexing systems', Aslib Proceedings, 421-431.

Clinchant, S., Csurka, G. and Ah-Pine, J. (2011) 'Semantic Combination of Textual and Visual Information in Multimedia Retrieval', Proceedings of the 1st ACM International Conference on Multimedia Retrieval, Trento, Italy, 44:1--44:8.

Clinchant, S., Csurka, G., Ah-Pine, J., Jacquet, G., Perronnin, F., Sánchez, J. and Minoukadeh, K. (2010) 'XRCE's Participation in Wikipedia Retrieval, Medical Image Modality Classification and Ad-hoc Retrieval Tasks of ImageCLEF 2010', ImageCLEF 2010 Working Notes.

Cluogh, P.D. and Sanderson, M. (2006) 'User experiments with the eurovision cross–language image retrieval system', *Journal of the American Society for Information Science and Technology*, pp. 679–708.

Csurka, G., Clinchant, S. and Popescu, A. (2011) 'XRCE's and CEA LIST's Participation at Wikipedia Retrieval of ImageCLEF 2011', in *CLEF 2011 working notes*.

Csurka, G., Dance, C., Fan, L., Willamowski, J. and Bray, C. (2004) 'Visual categorization with bags of keypoints', Workshop on Statistical Learning in Computer Vision, ECCV, 1--22.

Daroczy, B., Pethes, R. and Benczur, A.A. (2011) 'SZTAKI @ ImageCLEF 2011', in *CLEF 2011 working notes*.

Daroczy, B., Petras, I. and Benczur, A.A. (2010) 'SZTAKI @ ImageCLEF 2010', in *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy.

Datta, R., Joshi, S., Li, J. and Wang, J.Z. (2008) 'Image Retrieval: Ideas, Influences, and Trends of the New Age', *ACM Transactions on Computing Surveys*, vol. 40, no. 2.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009) 'ImageNet: A large-scale hierarchical image database', IEEE Conference on Computer Vision and Pattern Recognition, 2009 (CVPR 2009), 248 -255.

Depeursinge, A. and Muller, H. (2010) 'Fusion Techniques for Combining Textual and Visual Information Retrieval', in *Experimental Evaluation in Visual Information Retrieval*.

Deselaers, T. and Ferrari, V. (2011) 'Visual and semantic similarity in ImageNet', IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1777 -1784.

Díaz Galiano, M.C. (2011) *Tesis Doctoral: Recuperacion de informacion multimodal basada en integracion del conocimiento*, Jaen.

Donald, K. and Smeaton, A. (2005) 'A Comparison of Score, Rank and Probability-Based Fusion Methods for Video Shot Retrieval', in *Image and Video Retrieval*.

Douze, M., Guillaumin, M., Mensink, T., Schmid, C. and Verbeek, J. (2009) 'INRIA–LEARs participation to ImageCLEF 2009', in *Working Notes of CLEF 2009*, Corfu, Greece.

El Demerdash, O., Kosseim, L. and Bergler, S. (2009) 'Image retrieval by inter–media fusion and pseudo–relevance feedback', in *Evaluating systems for multilingual and multimodal information access*.

Escalante, H., Hernadez, C., Sucar, L.E. and Montes, M. (2008) 'Late fusion of heterogeneous methods for multimedia image retrieval', Proceedings of the 1st ACM international conference on Multimedia information retrieval, 172-179.

Escalante, H.J., Hernández, C.A., Sucar, L.E. and Montes, M. (2008) 'Late fusion of heterogeneous methods for multimedia image retrieval', Proceedings of the 1st ACM international conference on Multimedia information retrieval, Vancouver, British Columbia, Canada, 172--179.

Ferecatu, M. and Sahbi, H. (2008) 'TELECOM ParisTech at ImageClefphoto 2008: Bi–modal text and image retrieval with diversity enhancement', in *Working Notes of CLEF 2008*, Aarhus, Denmark.

Fernández Salido, J.M. and Murakami, S. (2003) 'Extending Yager's orness concept for the \

Filev, D.P. and Yager, R.R. (1997) 'Operations on fuzzy numbers via fuzzy reasoning', *Fuzzy Sets and Systems*, vol. 91, no. 2, pp. 137 - 142, Available: 0165-0114.

Fox, C. (1992) 'Lexical Analysis and Stoplist', in *Information Retrieval: Data, Structures and Algorithms*, Prentice Hall.

Fox, E.A. and Shaw, J.A. (1994) 'Combination of multiple searches', Text Retrieval Conference, 243–252.

Furnas, G.W., Deerwester, S., Dumais, S.T., Landauer, T.K., Harshman, R.A., Streeter, L.A. and Lochbaum, K.E. (1988) 'Information retrieval using a singular value decomposition model of latent semantic structure', Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval, Grenoble, France, 465-480.

Gao, S. and Lim, J.H. (2009) 'I2R at ImageCLEF photo retrieval 2009', in *Working Notes of CLEF 2009*, Corfu, Greece.

Garcia-Serrano, A., Benavent, X., Granados, R., de Ves, E. and Goñi-Menoyo, J.M. (2010) 'Multimedia Retrieval by Means of Merge of Results from Textual and Content Based Retrieval Subsystems', in *Multilingual Information Access Evaluation II. Multimedia Experiments*.

Garcia-Serrano, A., Benavent, X., Granados, R. and Goñi-Menoyo, J.M. (2009) 'Some results using different approaches to merge visual and text-based features in CLEF'08 photo collection', in *Lecture Notes in Computer Science, Evaluating Systems for Multilingual and Multimodal Information Access*.

Gospodnetic, O. and Hatcher, E. (2004) *Lucene in Action*, Manning Publications.

Granados Muñoz, R., García Serrano, A. and Goñi Menoyo, J. (2009) 'La herramienta IDRA (Indexing and Retrieving Automatically)', *Procesamiento de Lenguaje Natural*, vol. 1, no. 43, Available: 1989-7553.

Granados Muñoz, R., García-Serrano, A. and Goñi Menoyo, J.M. (2009) 'La herramienta IDRA (Indexing and Retrieving Automatically)', in *Actas del XXV Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN'09)*, Sociedad Española para el Procesamiento del Lenguaje Natural.

Granados, R., Benavent, J., Benavent, X., de Ves, E. and Garcia-Serrano, A. (2011) 'Multimodal Information Approaches for the Wikipedia Collection at ImageCLEF 2011', CLEF (Notebook Papers/Labs/Workshop).

Granados, R., García-Serrano, A., Méndez, N. and Benavent, X. (2012) 'Experimentación en la Búsqueda de Imágenes a partir de Características Visuales y Textuales: Fusión Tardía y Expansión de la Consulta', *Procesamiento del Lenguaje Natural*, no. 48, pp. 73-80, Available: 1135-5948.

Grubinger, M., Clough, P.D., Müller, H. and Deselaers, T. (2006) 'The IAPR Benchmark: A New Evaluation Resource for Visual Information Systems', International Conference on Language Resources and Evaluation, Genoa, Italy.

Hanbury, A., Muller, H. and Clough, P. (2010) 'Special issue on image and video retrieval evaluation', *Computer Vision and Image Understanding*, pp. 409 - 410, Available: 10.1016/j.cviu.2010.02.002, 1077-3142.

Hare, J.S., Dupplaw, D.P. and Lewis, P.H. (2009) 'IAM@ImageCLEFphoto 2009: Experiments on maximising diversity using image features', in *Working Notes of CLEF 2009*, Corfu, Greece.

Hearst, M.A. (2009) *Search user interfaces*, Cambridge University Press.

Hernández, D., Fresno, V., Granados, R. and García-Serrano, A. (2011) 'Evaluando Modelos de Indexación Multilingüe en Recuperación de Imágenes Basada en Texto', CAEPIA 2011, Tenerife, España.

Holzapfel, H., Nickel, K. and Stiefelhagen, R. (2004) 'Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3d pointing gestures', ACM International Conference on Multimedia Interfaces, 175-182.

Hua, X. and Zhang, H. (2005) 'An Attention-Based Decision Fusion Scheme for Multimedia Information Retrieval', in *Advances in Multimedia Information Processing - PCM 2004*.

Ingwersen, P. and Jrvelin, K. (2011) *The Turn: Integration of Information Seeking and Retrieval in Context*, Springer Publishing Company, Incorporated.

Inkpen, D., Stogaitis, M., DeGuire, F. and Alzghool, M. (2008) 'Clustering for photo retrieval at ImageCLEF 2008', in *Working Notes of CLEF 2008*, Aarhus, Denmark.

Inoue, M. and Grover, P. (2008) 'Effects of visual concept–based post–retrieval clustering in ImageCLEFphoto 2008', in *Working Notes of CLEF 2008*, Aarhus, Denmark.

Iyengar, G., Nock, H.J. and Neti, C. (2003) 'Audio-visual synchrony for detection of monologues in video archives', Proceedings. 2003 International Conference on Multimedia and Expo, 2003. ICME '03, 329-32.

Jain, A., Nandakumar, K. and Ross, A. (2005) 'Score normalization in multimodal biometric systems', *Pattern Recognition*, vol. 38, no. 12, pp. 2270 - 2285, Available: 0031-3203, 10.1016/j.patcog.2005.01.012.

Jaramillo, G.E. and Branch Bedoya, J.W. (2008) 'DORIS: Sistema para la Recuperación de Imágenes de Piezas Mecánicas y de Automoción utilizando Descriptores de Textura', *Avances en Sistemas e Informática*, vol. 5, no. 2, Available: 1657-7663.

Jeon, J., Lavrenko, V. and Manmatha, R. (2003) 'Automatic image annotation and retrieval using cross-media relevance models', Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, Toronto, Canada, 119--126.

Jones, K.S. (1981) 'The Cranfield tests', in Jones, K.S. *Information Retrieval Experiment*, Butterworths, London.

Kekre, H.B., Mishra, D. and Kariwala, A. (2011) 'Survey Of Cbir Techniques And Semantics', *International Journal of Engineering Science and Technology (IJEST)*, vol. 3, no. 5.

Kittler, J., Hatef, M., Duin, R.P.W. and Matas, J. (1998) 'On combining classifiers', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226 -239, Available: 0162-8828, 10.1109/34.667881.

Kludas, J., Bruno, E. and Marchand-Maillet, S. (2008) 'Information Fusion in Multimedia Information Retrieval', in *Adaptive Multimedial Retrieval: Retrieval, User, and Semantics*.

Lana-Serrano, S., Villena-Román, J. and González-Cristobal, J.C. (2010) 'DAEDALUS at ImageCLEF Wikipedia Retrieval 2010: Expanding with Semantic Information from Context', in *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy.

Lee, J.H. (1997) 'Analyses of multiple evidence combination', *SIGIR Forum*, pp. 267--276, Available: 10.1145/278459.258587.

León, T., Zucarello, P., Ayala, G., de Ves, E. and Domingo, J. (2007) 'Applying logistic regression to relevance feedback in image retrieval systems', *Pattern Recognition*, vol. 40, no. 10, pp. 2621 - 2632, Available: 10.1016/j.patcog.2007.02.002, 0031-3203.

Lestari, M., Sanderson, M. and Clough, P. (2010) 'Diversity in Photo Retrieval: Overview of the ImageCLEFPhoto Task 2009', in *Multilingual Information Access Evaluation II. Multimedia Experiments*, Springer.

Lew, M.S., Sebe, N., Djeraba, C. and Jain, R. (2006) 'Content-based multimedia information retrieval: State of the art and challenges', *ACM Transactions on Multimedia Computing, Communications and Applications (ACM TOMCCAP)*, pp. 1–19.

Liu, N., Dellandrea, E., Chen, L., Trus, A., Zhu, C., Zhang, Y., Bichot, C.E., Bres, S. and Tellez, B. (2012) 'LIRIS-Imagine at ImageCLEF 2012 Photo Annotation task', ImageCLEF 2012 Working Notes, Roma, Italy.

Magalhaes, J. and Ruger, S. (2007) 'Information-theoretic semantic multimedia indexing', Proceedings of the 6th ACM international conference on Image and video retrieval, Amsterdam, The Netherlands, 619--626.

Mahalanobis, P. (1936) 'On the generalised distance in statistics', Proceedings of the National Institute of Sciences, 55-79.

Maillot, N., Chevallet, J.-P. and Lim, J. (2006) 'Inter-media Pseudo-relevance Feedback Application to ImageCLEF 2006 Photo Retrieval', Evaluation of Multilingual and Multi-modal Information Retrieval, 735-738.

Martínez Méndez, F.J. (2004) *Recuperación de información: modelos, sistemas y evaluación*, Murcia.

Martinez, G. and Benavent, X. (2010) *Plataforma gráfica vía Web para trabajar en el Foro de recuperación visual de la información ImageCLEF*, Universidad de Valencia.

McCandless, M., Hatcher, E. and Gospodnetić, O. (2010) *Lucene in Action, Second Edition*.

Min, J., Laveling, J. and Jones, G. (2010) 'Document Expansion for Text-based Image Retrieval atWikipediaMM2010', in *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy.

Montague, M. and Aslam, J.A. (2001) 'Relevance score normalization for metasearch', Proceedings of the tenth international conference on Information and knowledge management, CIKM '01, New York, NY, USA, 427--433.

Montague, M. and Aslam, J.A. (2002) 'Condorcet fusion for improved retrieval', CIKM02: Proceedings of the 11th international conference on Information and knowledge management, New York, NY, USA, 538-548.

Moulin, C., Barat, C., Lemaitre, C., Gery, M., Ducottet, C. and Largeron, C. (2009) 'Combining text/image in WikipediaMM task 2009', in *Working Notes of CLEF 2009*, Corfu, Greece.

Mulhem, P., Chevallet, J.P., Quenot, G. and Al Batal, R. (2009) 'MRIM–LIG at ImageCLEF 2009: Photo retrieval and photo annotation tasks', in *Working Notes of CLEF 2009*, Corfu, Greece.

Muller, H., Clough, P. and Desealaers, T. (2010) *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*, Heidalberg: Springer.

Naphade, M., Smith, J.R., Tesic, J., Chang, S.F., Hsu, W., Kennedy, L., Hauptmann, A. and Curtis, J. (2006) 'Large-Scale Concept Ontology for Multimedia', *IEEE Multimedia Magazine*.

Navarro, S., Muñoz, R. and Llopis, F. (2009) 'Evaluating fusion techniques at different domains at Image-CLEF subtasks', in *Working Notes of CLEF 2009*, Corfu, Greece.

Nefian, A.V., Liang, L., Pi, X., Liu, X. and Murphye, K. (2002) 'Dynamic bayesian networks for audio-visual speech recognition', *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 1, pp. 1274--1288, Available: 10.1155/S1110865702206083, 1110-8657.

Nock, H.J., Iyengar, G. and Neti, C. (2002) 'Assessing face and speech consistency for monologue detection in video', Proceedings of the tenth ACM international conference on Multimedia, 303--306.

Oard, D.W. (1997) 'Cross-Language Text Retrieval Research in the USA', Proceedings of the 3rd ERCIM DELOS Workshop on Multilingual Information Retrieval, Zurich, Switzerland.

Perronnin, F. and Dance, C. (2007) 'Fisher Kernels on Visual Vocabularies for Image Categorization', IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07, 1 -8.

Pfleger, N. (2004) 'Context based multimodal fusion', Proceedings of the 6th international conference on Multimodal interfaces, 265-272.

Popescu, A. (2010) 'Telecom Bretagne at ImageCLEF WikipediaMM 2010', in *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy.

Popescu, A., Le Borgne, H. and Moellic, P.A. (2008) 'Conceptual image retrieval over the Wikipedia corpus', in *Working Notes of CLEF 2008*, Aarhus, Denmark.

Popescu, A., Tsikrika, T. and Kludas, J. (2010) 'Overview of the wikipedia retrieval task at ImageCLEF 2010', in *Working Notes of CLEF 2010*, Padova, Italy.

Porter, M.F. (1980) 'An algorithm for suffix stripping', *Program*, p. 130−137.

Porter, M.F. (1997) 'An algorithm for suffix stripping', in Sparck Jones, K. and Willett, P. *Readings in information retrieval*.

Radova, V. and Psutka, J. (1997) 'An approach to speaker identification using multiple classifiers', IEEE International Conference on Acoustics, Speech, and Signal Processing, 1135 -1138.

Reddy, B.S. (2007) *Evidential reasoning for multimodal fusion in human computer interaction*.

Ren, F. and Bracewell, D.B. (2009) 'Advanced Information Retrieval', *Electronic Notes in Theoretical Computer Science*, vol. 225, no. 0, pp. 303 - 317, Available: 1571-0661, 10.1016/j.entcs.2008.12.082.

Renda, M.E. and Straccia, U. (2003) 'Web metasearch: rank vs. score based rank aggregation methods', Proceedings of the 2003 ACM symposium on Applied computing, SAC '03, Melbourne, Florida, 841--846.

Robertson, S. (2008) 'On the history of evaluation in IR', *Journal of Information Science*, vol. 34, no. 4, pp. 439-456.

Robertson, S. (2008) 'On the history of evaluation in IR', *Journal of Information Science*, pp. 439-456.

Robertson, S.E. and Jones, K.S. (1976) 'Relevance weighting of search terms', *Journal of the American Society for Information Science*, Available: 1097-4571, 10.1002/asi.4630270302.

Robertson, S.E. and Jones, K.S. (1976) 'Relevance weighting of search terms', *Journal of the American Society for Information Science*, vol. 27, Available: 1097-4571.

Robertson, S., Walker, S., Hancock, M., Gull, A. and Lau, M. (1993) 'Okapi at TREC', The First Text REtrieval Conference (TREC-1), 21-30.

Rui, Y., Huang, T.S. and Chang, S.-F. (1999) 'ImageRetrieval: Current Techniques, Promising Directions, and Open Issues', *Journal of Visual Communication and Image Representation*, vol. 10, pp. 39-62, Available: http://dx.doi.org/10.1006/jvci.1999.0413.

Ruiz, M.E. (2009) 'UNT at ImageCLEFmed 2009', in *Working Notes of CLEF 2009*, Corfu, Greece.

Ruiz, M.E., Chen, J., Pusapathy, K., Chin, P. and Knudson, R. (2010) 'UNT at ImageCLEF 2010: CLIR for Wikipedia Images', in *Working notes of the ImageCLEF 2010 Lab*, Padua, Italy.

Ruiz, M.E., Leong, C.W. and Hassan, S. (2011) 'UNT at ImageCLEF 2011: Relevance Models and Salient Semantic Analysis for Image Retrieval', in *CLEF 2011 working notes*.

Salton, G. and Buckley, C. (1988) 'Term weighting approaches in automatic text retrieval', *Information Processing and Management*, pp. 513–523.

Salton, and McGill, M.J. (1986) *Introduction to Modern Information Retrieval*, New York, NY, USA: McGraw-Hill, Inc.

Salton, G., Wong, A. and Yang, C.S. (1975) 'A vector space model for automatic indexing', *Commun. ACM*, pp. 613--620, Available: 0001-0782, 10.1145/361219.361220.

Sanderson, C. and Paliwal, K.K. (2004) 'Identity verification using speech and face information', *Digital Signal Processing*, pp. 449 - 480.

Saracevic, T. (1995) 'Evaluation of evaluation in information retrieval', Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, United States, 138--146.

Schauble, P. (1997) 'Text Retrieval', in Schauble, P. *Multimedia Information Retrieval*, Springer US.

Seco, A., Markonis, D. and Muller, H. (2012) 'The medGIFT Group in ImageCLEFmed 2012', ImageCLEF 2012 Working Notes.

Simpson, M., Rahman, M.M., Demner-Fushman, D., Antani, S. and Thoma, G.R. (2009) 'Text– and content–based approaches to image retrieval for the ImageCLEF 2009 medical retrieval track', in *Working Notes of CLEF 2009*, Corfu, Greece.

Smeulders, A., Worring, M., Santini, S., Gupta, A. and Jain, R. (2000) 'Content–based image retrieval at the end of the early years', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1349–1380.

Smith, J.R. and Chang, S.-F. (1996) 'Automated binary texture feature sets for image retrieval', IEEE International Conference on Acoustics, Speech, and Signal Processing, 2239 - 2242.

Snoek, C.G.M. and Smeulders, A.W.M. (2010) 'Visual-Concept Search Solved?', *IEEE Computer*, vol. 43, no. 6, pp. 76-78, Available: 0018-9162.

Snoek, C.G.M., Worring, M. and Smeulders, A.W.M. (2005) 'Early versus late fusion in semantic video analysis', MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia, ACM, New York, NY, USA, 399–402.

Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.M. and Smeulders, A.W.M. (2006) 'The challenge problem for automated detection of 101 semantic concepts in multimedia', Proceedings of the 14th annual ACM international conference on Multimedia, 421--430.

Thompson, B. (1994) *The concept of statistical significance testing*, ERIC Clearinghouse.

Tsikrika, T. (2010) *Image Retrieval in CLEF*, [Online], Available: http://imageclef.org/2010/wiki.

Tsikrika, T. (2011) *Image Retrieval in CLEF*, [Online], Available: http://imageclef.org/2011/wikipedia.

Tsikrika, T. (2011) *ImageCLEF - Image Retrieval in CLEF*, [Online], Available: http://imageclef.org/2011/wikipedia.

Tsikrika, T., Herrera, A.S. and Muller, H. (2011) 'Assessing the Scholarly Impact of ImageCLEF', in *Multilingual and Multimodal Information Access Evaluation*, Springer Berlin Heidelberg.

Tsikrika, T., Popescu, A. and Kludas, J. (2011) 'Overview of the Wikipedia Image Retrieval Task at ImageCLEF 2011', CLEF (Notebook Papers/Labs/Workshop).

van Rijsbergen, C.J. (1999) *Information Retrieval*, Glasgow.

van Rijsbergen, C.J., Robertson, S.E. and Porter, M.F. (1980) *New models in probabilistic information retrieval*, London.

Vanegas, J.A., Caicedo, J.C., Camargo, J.E., Ramos-Pollan, R. and Gonzalez, F.A. (2012) 'Bioingenium at ImageCLEF 2012: Textual and Visual Indexing for Medical Images', ImaheCLEF 2012 Working Notes, Roma, Italy.

Villena-Roman, J., Lana-Serrano, S., Martinez-Fernandez, J.L. and Gonzalez-Cristobal, J.C. (2007) 'MIRACLE at ImageCLEFphoto 2007: Evaluation of merging strategies for multilingual and multimedia information retrieval', in *Working Notes of CLEF 2007*, Budapest, Hungary.

Vogt, C.C. and Cottrell, G.W. (1999) 'Fusion Via a Linear Combination of Scores', *Information Retrieval*, vol. 1, no. 3, pp. 151-173, Available: 1386-4564, 10.1023/A:1009980820262.

Voorhees, E. (2006) 'Common Evaluation Measures', Proceedings of Text Retrieval Conference (TREC-15).

Wang, J., Kankanhalli, M., Yan, W. and Jain, R. (2003) 'Experiential Sampling for video surveillance', First ACM SIGMM international workshop on Video surveillance, Berkeley, California, 77--86.

Wang, J., Song, D. and Kaliciak, L. (2010) 'RGU at ImageCLEF2010 Wikipedia Retrieval Task', in *CLEF 2010 LABs and Workshops, Notebook Papers*, Padua, Italy.

Wong, S.K.M., Ziarko, W. and Wong, P.C.N. (1985) 'Generalized vector spaces model in information retrieval', Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval, Montreal, Quebec, Canada, 18-25.

Wu, H. (2003) *Sensor data fusion for context-aware computing using dempster–shafer theory*.

Wu, Z., Cai, L. and Meng, H. (2006) 'Multi-level fusion of audio and visual features for speaker identification', International Conference on Advances in Biometrics, 493–499.

Wu, Y., Chang, E. and Tsengh, B.L. (2005) 'Multimodal metadata fusion using causal strength', ACM International Conference on Multimedia, Singapore, 872–881.

Wu, S., Crestani, F. and Bi, Y. (2006) 'Evaluating score normalization methods in data fusion', in *Information Retrieval Technology, AIRS 2006*.

Yager, R. (1988) 'On ordered weighted averaging aggregation operators in multi criteria decision making', *IEEE Transaction Systems Man and Cybernetics*, pp. 183-190.

Yan, R. (2006) *Probabilistic models for combining diverse knowledge sources in multimedia retrieval*.

Yan, X., Wu, W., Gao, G. and Lu, Q. (2012) 'IMU @ ImageCLEF 2012', ImageCLEF 2012 Working Notes.

Zagoris, K., Arampatzis, A. and Chatzichristofis, S.A. (2010) 'www.MMRetrieval.net: a multimodal search engine', in *Proceedings of the Third International Conference on SImilarity Search and APplications, SISAP '10*, New York, NY, USA: ACM.

Zhou, X., Depeursinge, A. and Müller, H. (2010) 'Information Fusion for Combining Visual and Textual Image Retrieval', International Conference on Pattern Recognition.

Zhou, X., Gobeill, J. and Muller, H. (2008) 'MedGIFT at ImageCLEF 2008', in *Working Notes of CLEF 2008*, Aarhus, Denmark.

Zhu, Q., Yeh, M.C. and Cheng, K.T. (2006) 'Multimodal fusion using learned text concepts for image categorization', ACM International Conference on Multimedia, 211–220.

Znaidia, A., Shabou, A., Popescu, A. and Le Borgne, H. (2012) 'CEA LIST's participation to the Concept Annotation Task of ImageCLEF 2012', ImageCLEF 2012 Working Notes, Roma, Italy.

Zubiaga, A. (2011) *PhD thesis: Harnessing Folksonomies for Resource Classification*, Madrid, Spain.