Estudio sobre el impacto de los componentes de un sistema de recuperación de información geográfica y temporal

Fernando S. Peregrino
Universidad de Alicante
fsperegrino@dlsi.ua.es

David Tomás Díaz
Universidad de Alicante
dtomas@dlsi.ua.es

Fernando Llopis Pascual
Universidad de Alicante
llopis@dlsi.ua.es

Resumen

La inmensa mayoría de los motores de búsqueda comerciales se centran principalmente en la recuperación de información textual, tratando de igual forma cualquier otro tipo de información contenida en el texto. Dicho tratamiento hace que cuando se añade alguna otra dimensión, como pueden ser la geográfica o la temporal, los citados buscadores obtienen unos pobres resultados. El presente trabajo pretende centrarse en los sistemas de recuperación de información geográfica y temporal abordando toda la problemática relacionada con éstos. Para ello, se ha desarrollado un sistema completo para el tratamiento de la dimensión geográfica y temporal en el texto y su aplicación a la recuperación de información, basando el citado sistema en múltiples motores de búsqueda y en técnicas de búsqueda de respuestas. Este sistema se ha evaluado en la tarea *GeoTime* del *workshop NTCIR*, lo que ha permitido comparar el sistema con otras aproximaciones actuales al tema.

Palabras clave: recuperación de información geográfica, etiquetado geográfico, información espacial, información temporal.

1. Introducción

En la sociedad actual, prácticamente se puede acceder a toda la información en formato digital. Dicha información se encuentra en constante crecimiento, por lo que se hacen imprescindibles herramientas que sean capaces de obtener los documentos deseados de una forma eficaz, rápida y sencilla.

La recuperación de la información (*IR: Information Retrieval*) es la ciencia de la búsqueda de información en documentos electrónicos dando como resultado un conjunto de estos mismos documentos ordenados según la relevancia que tengan con la consulta formulada.

Según un estudio realizado por (Zhang, Rey, y Jones, 2006), el 12,7 % sobre 4 millones de consultas de ejemplo contenía un topónimo, de lo que se desprende que la geografía también se ve involucrada en *IR*. Consultas del tipo *"Catedrales en Europa"*, o *"Dónde murió Osama bin Laden"*, hacen necesaria la intervención de dicha materia.

La recuperación de información geográfica (*GIR: Geographical Information Retrieval*) es una especialización de *IR* con metadatos geográficos asociados. Los sistemas de *IR*, generalmente, ven los documentos como una colección o "bolsa de palabras". Por el contrario, los sistemas *GIR* necesitan información semántica, es decir, necesitan de un lugar o rasgo geográfico asociado a un documento. Debido a esto, es común en los sistemas *GIR* que se separe el análisis y la indexación de texto de la indexación geográfica.

En la recuperación de información sobre textos sin estructurar se dificulta aún más la obtención del topónimo al que se hace referencia. La estructura típica de una consulta que requiere de *GIR* es <qué_se_busca> + <relación> + <localización>. Si nos centramos en un ejemplo concreto, dada la siguiente consulta: "*Estaciones de esquí en España*", deberíamos de limitar los resultados devueltos a aquellas estaciones ubicadas dentro del ámbito geográfico de la consulta ("*España*").

Los sistemas *GIR* son un campo de investigación en auge en los últimos años debido a la falta de buenos resultados cuando se realiza una búsqueda centrada en una ubicación específica. Son diversas las competiciones que se han organizado alrededor de este tipo de sistemas. El *CLEF*[1] (*Cross Language Evaluation Forum*) agregó una rama geográfica, *GeoCLEF*[2]. A raíz de esta rama geográfica, nacieron otro tipos de tareas como el *GikiP*, la cual fue una tarea piloto en el *GeoCLEF* 2008 pasando en 2009 a llamarse *GikiCLEF* y ser una tarea propia dentro del *CLEF*. Ésta tarea consistía en encontrar entradas

[1] http://www.clef-campaign.org/
[2] http://ir.shef.ac.uk/geoclef/

o documentos en la *Wikipedia* que contestaran a una serie de consultas que requerían de algún tipo de razonamiento geográfico. El *NTCIR*[3] (*NII Test Collection for IR Systems*) creó la tarea *GeoTime*[4]. Esta tarea combina *GIR* con búsqueda basada en el tiempo para encontrar eventos específicos en una colección de textos. También se han creado *workshops* específicos en la materia como el *Geographic Information Retrieval*[5].

Este artículo pone su foco de atención en los sistemas *GIR*, realizando un estudio exhaustivo de la situación actual en dicha materia. Además del presente estudio, se ha desarrollado un sistema *GIR* modular con el fin de discutir las dificultades expuestas en éste, evaluado cómo afectan los distintos componentes que intervienen en el proceso sobre el rendimiento final del sistema. Para dicha empresa, se ha evaluado el sistema en la tarea *GeoTime* del *NTCIR*, para lo que se ha tenido que incorporar al sistema un módulo para el tratamiento de la información temporal.

Este trabajo está estructurado según sigue. Primero, se introduce el estado de la cuestión que recopila los trabajos relacionados más relevantes hasta la fecha, así como las principales tendencias en este campo. A continuación, se da paso a la descripción detallada del sistema implementado deteniéndonos en cada uno de sus módulos y componentes. Se prosigue con los experimentos realizados en dicho sistema, así como la evaluación obtenida de los mismos y un análisis exhaustivo de los resultados. Para finalizar, se muestran las conclusiones y trabajo futuro para extender y mejorar el sistema.

2. Trabajo relacionado

En los últimos años ha habido un incremento en la investigación dedicada a la recuperación de información geográfica dado su gran interés mercantil. Los grandes motores de búsqueda web comerciales (*Google*, *Yahoo!* y *Bing*) han desarrollado herramientas para poder afrontar dicha problemática, sin embargo, dichas herramientas tienen un amplio margen de mejora.

Si hay un proyecto que es referencia obligatoria para todo aquel que se quiera iniciar en la materia, y que aún hoy en día sigue marcando la pauta a seguir por el resto de investigadores en *GIR*, ese es el proyecto *SPIRIT*(Jones et al., 2007). En este proyecto se crearon herramientas software y técnicas que pueden ser usadas para crear motores de búsqueda y sitios web que muestren inteligencia en el reconocimiento de terminología geográfica. Con el fin de demostrar y evaluar los resultados del proyecto, se construyó un prototipo de motor de búsqueda *GIR*, el cual está siendo usado como plataforma para probar y evaluar nuevas técnicas en recuperación de información geográfica. Este proyecto aborda plenamente todos los frentes abiertos en la investigación en *GIR*.

En (Wang et al., 2005) clasificaron las áreas de investigación en esta materia en tres grandes grupos: identificación y desambiguación de topónimos (etiquetado geográfico), desarrollo de herramientas informáticas para el manejo de la información geográfica, y explotación de las diversas fuentes de recursos geográficos. En (Jones y Purves, 2008) podemos ver una disección más exhaustiva de los principales asuntos que podemos abordar en los sistemas *GIR*:

1. Detección de referencias geográficas.
2. Desambiguación de topónimos.
3. Terminología geográfica vaga.
4. Indexación espacial y geográfica.
5. Ranking por relevancia geográfica.
6. Iterfaces de usuario.
7. Métodos de evaluación de los sistemas *GIR*.

A continuación, se van a ver las distintas aproximaciones que se han llevado a cabo en cada uno de estos apartados.

2.1. Detección de referencias geográficas

La detección de referencias geográficas o *geotagging* en el Procesamiento del Lenguaje Natural (*PLN*) es una extensión de los reconocedores de entidades nombradas (*NER: Named Entity Recognition*), y versa sobre el análisis de los textos con el fin de identificar la presencia inequívoca de topónimos. Dicha problemática dista mucho de ser algo trivial, dado que en numerosas ocasiones podemos encontrarnos con que los nombres de personas, organizaciones, equipos deportivos, etc., son compartidos por los topónimos. Podemos ver como (Li et al., 2006) afronta la tarea con un mecanismo para la resolución probabilística de topónimos, dando dicho método una mejora de la efectividad sobre el subconjunto de *topics* del *GeoCLEF*.

Otra dificultad con la que nos topamos es la metonimia, en situaciones tales como "*no aceptaremos órdenes de Madrid*". En (Leveling y Hartrumpf, 2007) se muestra un enfoque para solucionar el problema de la metonimia a través del análisis de rasgos superficiales.

[3] `http://research.nii.ac.jp/ntcir/index-en.html`
[4] `http://metadata.berkeley.edu/NTCIR-GeoTime/`
[5] `http://www.geo.uzh.ch/~rsp/gir10/`

2.2. Desambiguación de topónimos

Una vez obtenido de manera inequívoca el nombre del lugar hay que desambiguarlo, ya que, muchos de los topónimos existentes son compartidos por varios lugares (p. ej. Granada, Springfield, etc.). Para dicha tarea se han afrontado diversas estrategias tales como la identificación por medio del resto de topónimos del texto, es decir, obteniendo el ámbito del que se está hablando para desambiguar cada uno de los lugares (Wang et al., 2005). Otra de las estrategias seguidas en esta materia, al hilo de la anterior, es el esclarecimiento del lugar nombrado jugando con las entidades geográficas de orden superior o inferior mencionadas en el texto (Silva et al., 2006).

2.3. Terminología geográfica difusa

Otra problemática adicional es la de las expresiones geográficas difusas, es decir, aquellas que describen lugares imprecisos que no podemos encontrar en ninguna base de datos geográfica (*gazetteer*). Ampliando el ejemplo visto en la introducción, serían expresiones del tipo "*Estaciones de esquí en el norte de España*", donde la entidad geográfica en la que deberíamos buscar ("*norte de España*") resulta imprecisa.

Los trabajos previos realizados en el campo de la definición automática de regiones geográficas difusas han seguido dos líneas fundamentales a la hora de obtener la información necesaria para definir dichas regiones. La primera aproximación se centra en la obtención de información a partir de la consulta directa a un conjunto de usuarios reales, que son los encargados de delimitar la región a estudio. La segunda aproximación está enfocada a la obtención de información a partir de fuentes de información no estructurada para la identificación de estas regiones.

Dentro del primer grupo se encuentra el trabajo de (Montello et al., 2003). En este trabajo se propone una aproximación probabilística basada en frecuencias, donde la inclusión de una determinada localización dentro de una región difusa viene condicionada por el número de usuarios que considera su pertenencia a la misma. Estos valores sirven para generar una curva de nivel que delimita la región difusa y proporciona una probabilidad de pertenencia a las localizaciones que se hallan en su interior.

El objetivo de la segunda aproximación es recuperar suficiente información de la web para poder definir espacialmente las regiones difusas estudiadas. En esta línea, en (Clough, 2005) se creó un sistema donde las coordenadas de las localidades contenidas en dichas regiones se empleaban para la definición de polígonos representativos de dicha región. Una extensión de este trabajo se puede encontrar en (Jones et al., 2008), donde utilizan esta aproximación para identificar tanto regiones difusas como precisas.

2.4. Indexación espacial y textual

Existen una serie de técnicas para la indexación textual de documentos. Dichas técnicas, tal y como podemos ver en (Baeza-Yates y Ribeiro-Neto, 1999), suelen basarse en la creación de índices inversos, es decir, se añaden al índice todas las palabras encontradas en el corpus, indicando en qué documentos aparece cada palabra y con qué frecuencia, para su posterior recuperación mediante la intersección con la consulta del usuario. En cuanto a la indexación espacial, los sistemas de información geográfica (*GIS: Geographical Information System*) son los que han tratado con más éxito dicho asunto. La dificultad subyace en conseguir mezclar ambos índices con acierto (Cardoso y Santos, 2007). La técnica más común para afrontar dicha mezcla ha sido la obtención de "huellas" (*footprints*) espaciales de los documentos, para posteriormente poder indexar esas huellas por documentos. El problema es que un alto número de huellas por documento (según (Vaid et al., 2005) hay unas 21 por documento web) puede hacer intratable el problema, por lo que han habido muchos trabajos orientados a obtener un mínimo número de huellas representativas por documento (Wang et al., 2005)(Silva et al., 2006).

En (Vaid et al., 2005) podemos ver tres estilos diferentes de conseguir llevar a cabo dicha tarea con celeridad y mejorando los resultados de un sistema de *IR* genérico. En dicho trabajo se comparó la indexación textual corriente (*PT: Pure Text*), contra tres tipos de indexación textual y espacial: indexación texto-espacial (*TS: Text-primary spatio-textual indexing*), indexación espacio-textual (*ST: Space-primary spatio-textual indexing*) e indexación separada (*T: Separate text and spatial indexes*), dando como resultado una gran mejora en el *recall*, aunque las indexación *TS* y *ST* supusieron un considerable aumento en el espacio requerido para su indexación.

2.5. Ranking por relevancia geográfica

La clasificación de documentos relevantes o ranking, determina la manera en la que debe de puntuar un documento según su idoneidad para con la consulta lanzada para su posterior devolución al usuario. La manera más usual con la que suelen afrontar la clasificación de documentos los

principales motores de búsqueda es mediante la intersección de los términos de la consulta con los de los documentos, dando un mayor valor a aquellos términos que ocurran en un menor número de documentos, a los términos que con más frecuencia aparezcan en un documento, y a la proporción de apariciones de un término en un documento dada la longitud de éste último (Robertson, Walker, y Hancock-Beaulieu, 1998).

En el caso de los sistemas *GIR* no hay que devolver una simple intersección de términos, si no que hay que saber tratar semánticamente los topónimos referenciados en la consulta de tal forma que se capturen las huellas existentes en ésta y se comparen con las de los documentos del corpus, teniendo en cuenta que dichas huellas pueden pertenecer a un ámbito geográfico inferior, superior, intersectante, etc. (Van Kreveld et al., 2005).

2.6. Iterfaces de usuario

El desarrollo de una interfaz de usuario eficaz para los sistemas *GIR* es un tema que se ha tratado con escaso éxito dada su dificultad. La mayoría de consultas geográficas, como ya se ha indicado anteriormente, son del modo <qué_se_busca> + <relación> + <localización>, por lo que parece sencillo crear una interfaz con tres campos, pero habría que saber tratar cualquier tipo de "relación", y también habría que tener en cuenta que no todas las consultas son del tipo descrito anteriormente (p. ej. "*Dónde murió Osama bin Ladem*"). Otra aproximación que se ha hecho en (Jones et al., 2007) es permitir al usuario trazar una zona en un mapa en la que se quiere obtener los resultados e introducir la consulta en sí.

Por otro lado, nos encontramos con la problemática de cómo devolver los resultados. ¿Se debe adjuntar un mapa sobre los sitios de los que habla cada documento? ¿A qué escala? ¿Un mapa por documento o un único mapa para todos los documentos relevantes?

2.7. Métodos de evaluación de los sistemas *GIR*

La evaluación de los resultados devueltos por un sistema *GIR*, al igual que en los sistemas de *IR*, es una tarea costosa dado que hay que ir examinando manualmente, uno a uno, todos los documentos devueltos por un sistema para comprobar la relevancia de cada una de las preguntas que se le ha lanzado. En tareas como la del *GeoCLEF* o *NTCIR GeoTime*, dichas evaluaciones se han resuelto mediante valores binarios, es decir, diciendo si un documento es relevante o no para la pregunta realizada, aunque también es cierto que en el *NTCIR GeoTime*, dicho juicio se hizo añadiendo una mayor escala de valores, según lo relevante que sea. En *GIR* hay que tener en cuenta que se añade una nueva dimensión a evaluar, la geográfica, por lo que un documento puede ser relevante en cierto grado geográficamente hablando, y en otro grado en cuanto a contenido.

Por ejemplo, en la tarea *GeoTime* del *NTCIR 2011*, se emplearon tres métricas distintas para la evaluación de los documentos: *Average Precision* (*AP*), *Q-measure* (*Q*), y *normalised Discounted Cumulative Gain* (*nDCG*) (Mitamura et al., 2008). La evaluación se efectuó recogiendo los n primeros documentos que cada uno de los sistemas de los participantes entendió que eran relevantes para cada una de las consultas efectuadas. Dichos documentos se evaluaban de la siguiente manera:

- Relevante. Si contestaba a la pregunta dónde y cuándo.
- Parcialmente relevante (dónde). Si contestaba a la pregunta dónde.
- Parcialmente relevante (cuándo). Si contestaba a la pregunta cuándo.
- Parcialmente relevante (otro). Si contestaba de alguna manera a la pregunta.
- Irrelevante. Si no contestaba a la pregunta.

3. Descripción del sistema

Los sistemas *GIR*, comúnmente, se pueden dividir en los siguientes módulos: etiquetado geográfico (obtención y desambiguación de los topónimos para su posterior procesamiento), indexación geográfica y de texto, almacenamiento de datos, clasificación geográfica por relevancia (con respecto a una consulta geográfica) y la navegación en los resultados (normalmente con un mapa como interfaz).

Para la creación del presente sistema *GIR*, se ha optado por una implementación modular con el fin de poder así añadir nuevos componentes en un futuro y crear mejoras sobrelos ya existentes.

En la figura 1 se muestran en la parte izquierda y central los elementos que intervienen en la fase inicial de indexación y preprocesamiento del corpus. En la parte central y derecha se muestran los que intervienen en las fases que tienen que ver con las consultas, es decir, los de procesado y ejecución de las mismas. También se puede observar que las líneas continuas que unen los distintos componentes del sistema están relacionadas con todas las acciones de preprocesado, las líneas discontinuas indican las acciones llevadas

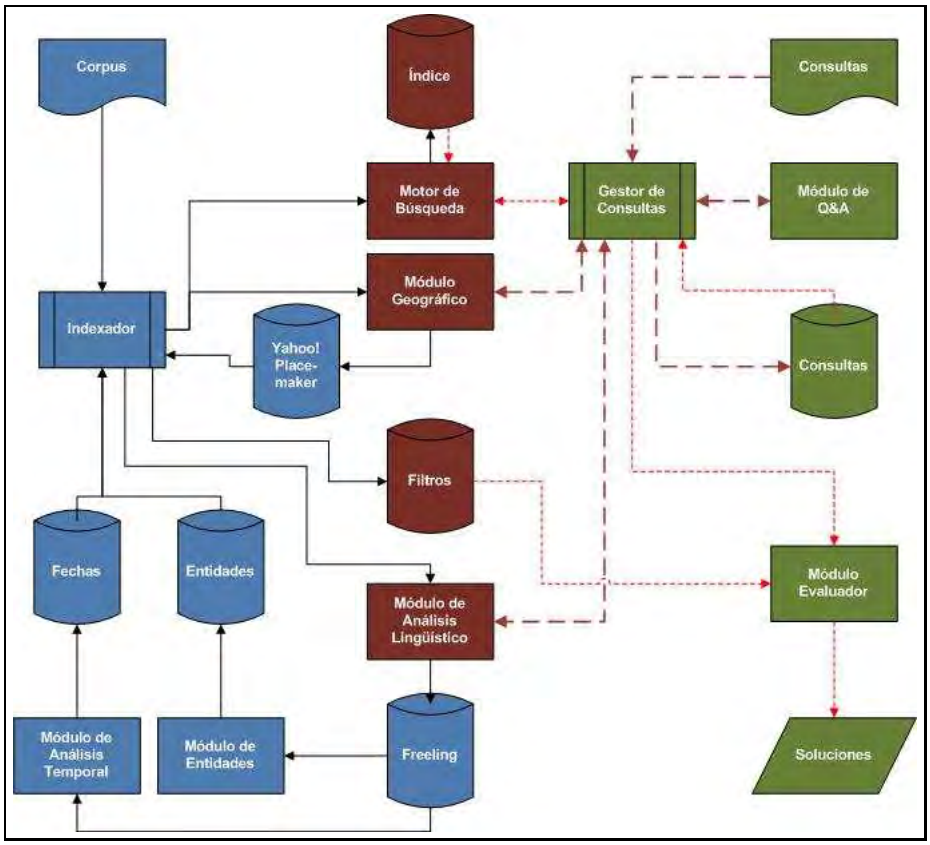


Figura 1: Esquema de funcionamiento general del sistema *GIR* implementado.

a cabo en tiempo de ejecución, siendo las líneas gruesas discontinuas las que indican el procesado de la consulta, mientras que las líneas discontinuas representan la búsqueda de la consulta en sí (sección 3.3).

Dado que se ha decidido evaluar el sistema construido en la tarea *GeoTime* del *NTCIR*, la cual tiene una componente temporal (tal y como se explicará más en profundidad en la sección 4), ha sido necesaria la creación de un módulo que sea capaz de tratar la información temporal.

3.1. Módulos del sistema

A continuación se describen en detalle cada uno de los módulos desarrollados en el sistema.

3.1.1. Motor de Búsqueda.

Como motor de búsqueda para este sistema, originalmente se optó por utilizar *Lucene*[6], un motor de búsqueda de alto rendimiento equipado con librerías de búsqueda escritas completamente en Java. Se trata de una tecnología adecuada para casi cualquier aplicación que requiera búsqueda de texto completo, especialmente en distintas plataformas. *Apache Lucene* es un proyecto de código abierto disponible para su descarga gratuita.

Viendo que la cobertura obtenida por *Lucene* apenas llegaba al 50 %, como se podrá ver con más detenimiento en la sección 4, y basándonos en el trabajo realizado por (Perea-Ortega, 2010), se decidió dotar al sistema de un motor de búsqueda adicional, *Terrier*[7]. Así pues, para obtener una puntuación definitiva normalizada por cada uno de los documentos devueltos por ambos motores de búsqueda, se procedió según sigue para cada una de las consultas lanzadas:

1. Se obtiene el valor máximo de *Lucene* de entre todos los documentos devueltos por éste.
2. Se dividen todas las puntuaciones de todos los documentos devueltos por *Lucene* entre el valor máximos indicado en el punto anterior.
3. Análogamente, se repiten los dos pasos anteriores para *Terrier*.
4. En caso de que haya documentos devueltos tanto por *Lucene* como por *Terrier*, se suman las puntuaciones de ambos.
5. Finalmente, se divide entre dos la puntuación de cada uno de los documentos devueltos por los motores de búsqueda citados an-

[6] `http://lucene.apache.org/`

[7] `http://terrier.org/`

teriormente, obteniendo así un valor normalizado entre 0 y 1.

El módulo del motor de búsqueda tiene principalmente dos funcionalidades: la indexación de todo el corpus y la recuperación de una serie de documentos dada una consulta.

Al motor de búsqueda se le han añadido una serie de características para mejorar su rendimiento, tales como un lematizador y la eliminación de palabras de parada (*stopwords*). Para la indexación del corpus, se obvian todas las *stopwords* y se indexan el lema del resto de palabras de cada uno de los documentos.

Para la ordenación de los resultados según su relevancia, se ha utilizado la función de pesado *BM25* (Robertson, Walker, y Hancock-Beaulieu, 1998) basada en los modelos probabilísticos de *IR*.

Por otro lado, también se ha utilizado el modelo de expansión de consulta *Bose-Einstein* (*Bo1*).

Finalmente, se ha establecido que el motor de búsqueda pueda recuperar hasta 1.000 documentos relevantes.

3.1.2. Módulo de Análisis Lingüístico.

Este módulo se encarga de:

- Lematizar la consulta.
- Eliminar las *stopwords*.
- Obtener los topónimos de las consultas mediante una base de datos de nombres de lugar (*GeoNames*[8]).
- Obtener las fechas de la consulta mediante un analizador lingüístico (*FreeLing*[9]).
- Obtener las restricciones geográficas y temporales de las consultas (p. ej. si en la búsqueda es suficiente con encontrar el nombre del país o hay que buscar el de la ciudad también, si con el año y el mes es suficiente o hay que encontrar la fecha completa también, etc.). Para ello se analizan las partes narrativas de las consultas en busca de palabras clave como "país", "estado", "ciudad", "día", "fecha exacta", etc.
- Encontrar otro tipo de entidades (p. ej. nombres de persona, nombres de empresas, etc.) mediante un analizador sintáctico (*FreeLing*).
- Obtener otros términos lingüísticos comunes para una posible expansión de la consulta.

[8] `http://www.geonames.org/`

[9] `http://nlp.lsi.upc.edu/freeling/`

3.1.3. Módulo de Q&A.

Una de las novedades que se ha introducido en este sistema *GIR* ha sido un módulo de búsqueda de respuesta (*Q&A: Question Aswering*), mediante el cual se pretenden obtener los términos geográficos y temporal que más se adecúan a la respuesta de la pregunta para su posterior intersección con los artículos del corpus. Lo que se hace en este módulo es lanzar la consulta a la interfaz de búsqueda de *Yahoo!* (*Yahoo! Search BOSS*[10]) y recoger los resúmenes de los 1000 primeros resultados devueltos. De estos resúmenes se extraen todas las fechas y lugares que se encuentren, se cuenta el número total de ocurrencias de cada uno de ellos y se normaliza, quedándose el sistema con los 10 más relevantes para cada una de las siguientes 4 categorías: fecha completa, mes y año, año, y topónimos.

3.1.4. Módulo de Análisis Temporal.

Como ya se ha indicado con anterioridad, este módulo ha sido introducido con el fin de poder evaluar el sistema en la tarea *GeoTime* del *NTCIR*, pero no forma parte del principal propósito de esta investigación. La implementación se ha apoyado en el analizador lingüístico *FreeLing*. Concretamente, se ha aprovechado que *FreeLing* dispone de un módulo de detección y normalización de fechas para obtener todas las referencias a fechas que haya en cada uno de los documentos del corpus, incluida la fecha de cada documento (la de publicación en el periódico, ya que los documentos fueron extraídos de noticias periodísticas). De esta forma, en el tiempo de preproceso se crea un nuevo fichero por cada artículo existente en el corpus, en el que están registradas todas las fechas que *FreeLing* ha detectado en el artículo. Posteriormente, en tiempo de ejecución, se busca intersectar la fecha que pueda haber en la consulta con la de los documentos devueltos por el motor de búsqueda, dando mayor peso si coincide la fecha completa (día, mes y año) a si lo hace parcialmente (mes y año, o solamente el año).

3.1.5. Módulo Geográfico.

En este módulo es donde se hará todo el tratamiento geográfico. Para el desarrollo de dicho módulo, el sistema se ha basado en *Yahoo! Placemaker*[11].

Yahoo! Placemaker es un servicio web de *geoparsing*[12] de libre disposición. Es útil para desarrolladores que quieren hacer aplicaciones basa-

[10] `http://developer.yahoo.com/search/boss/`

[11] `http://developer.yahoo.com/geo/placemaker/`

[12] Detección y desambiguación de nombres de lugar asignándole un identificador único.

das en localización espacial mediante la identificación de topónimos existentes en textos no estructurados (p. ej. *feeds*, páginas web, noticias, etc.), de los que es capaz de devolver metadatos geográficos asociados a dichos textos. La aplicación identifica los topónimos en el texto, los desambigua, y devuelve identificadores de lugar únicos para cada uno de los lugares que tiene en su base de datos. También aporta otro tipo de información como, cuántas veces aparece el lugar en el texto, en qué lugar del texto se encontró, etc.

En el caso concreto de este sistema *GIR*, ha utilizado *Yahoo! Placemaker* para la obtención de topónimos, la desambiguación de los mismos, y la obtención de entidades administrativas de orden superior e inferior.

Yahoo! Placemaker devuelve un documento *XML* por cada texto que se le pase. Finalmente, el módulo almacena toda la información geográfica pertinente del artículo en un documento *XML*.

Este módulo tiene otra función a la hora de analizar las consultas. Concretamente, lo que hace es recoger los topónimos existentes en los ficheros *XML* de las consultas y los transforma al identificador inequívoco de *Yahoo!* (*WOEID: Where On Earth IDentifier*) para un procesamiento más ágil en la fase de búsqueda.

3.1.6. Módulo para la Detección de Entidades.

Este módulo se encarga de guardar un documento por cada uno de los artículos del corpus. En dicho documento estarán todas las entidades reconocidas por *FreeLing* en el artículo original para, posteriormente, crear un nuevo documento *XML* que sirva de filtro para las consultas introducidas.

3.1.7. Módulo de Reordenación.

Este módulo entra en acción únicamente en tiempo de ejecución, es decir, en el momento de realizar la búsqueda de una consulta concreta. El objetivo de este módulo es intersectar los documentos *XML* de las consultas y los documentos *XML* de los artículos de corpus que han sido devueltos como solución por el motor de búsqueda. Como resultado de dicha intersección, el módulo evaluador reorganizará el ranking de documentos devuelto por el motor de búsqueda. Dicha reorganización viene dada por dos esquemas de pesado. En ambos esquemas de pesado se puede observar la importancia del peso de *Lucene*, pero se diferencian en:

- Esquema de pesado A. Permite evaluar la importancia del módulo de Q&A de la parte descriptiva (parte de la consulta en sí misma) y la parte narrativa (parte que detalla qué es lo que se busca concretamente, p. ej. ciudad, estado, fecha exacta, etc.) de la consulta:

$$\alpha\cdot(\beta\cdot\omega_{desc}+(1-\beta)\cdot\omega_{narr})+(1-\alpha)\cdot\omega_{QA} \quad (1)$$

  Donde:

  - α = Peso que se le da a la parte descriptiva y narrativa de la consulta.
  - β = Peso que se le da a la parte descriptiva de la consulta.
  - ω_{desc} = Valor normalizado de la parte descriptiva de la consulta.
  - ω_{narr} = Valor normalizado de la parte narrativa de la consulta.
  - ω_{QA} = Valor normalizado del módulo de Q&A.
- Esquema de pesado B. Permite evaluar la importancia de la parte geográfica, la temporal y la de entidades:

$$\alpha\cdot(\beta\cdot\omega_{geo}+(1-\beta)\cdot\omega_{temp})+(1-\alpha)\cdot\omega_{ent} \quad (2)$$

  Donde:

  - α = Peso que se le da a la parte geográfica y temporal de la consulta con respecto a la de entidades.
  - β = Peso que se le da a la parte geográfica de la consulta con respecto a la temporal.
  - ω_{geo} = Valor normalizado de la parte geográfica.
  - ω_{temp} = Valor normalizado de la parte temporal.
  - ω_{ent} = Valor normalizado de la parte de entidades.

En ambos esquemas de pesado, una vez obtenido el resultado de las ecuaciones 1 y 2, se tiene que unir con el resultado obtenido por el motor de búsqueda, utilizándose para ambos esquemas la siguiente fórmula:

$$\lambda \cdot L + (1 - \lambda) \cdot E \quad (3)$$

Donde:

- λ = Peso que se le da al motor de búsqueda (*Lucene+Terrier*).
- L = Valor normalizado de los resultados devueltos por el módulo de motor de búsqueda.
- E = Esquema de pesado elegido (ver ecuación 1 y 2).

3.2. Esquemas de Almacenamiento

Con el fin de agilizar y de hacer más eficiente el proceso a la hora de realizar consultas, se han guardado dos grupos de documentos *XML*: los documentos de filtrado por cada uno de los artículos del corpus y los documentos de análisis de cada una de las consultas.

3.2.1. Corpus.

Estos documentos *XML* se dividen en tres partes: geográfica, temporal y de entidades (figura 2).

```
▼<document id="NYT_ENG_20040502.0019">
 ▼<geo>
  ▼<entity>
     <documentType>ancestor</documentType>
     <woeid>23424977</woeid>
     <type>Country</type>
     <name>United States</name>
   </entity>
  ▼<entity>
     <documentType>geographicScope</documentType>
     <woeid>24701772</woeid>
     <type>Zone</type>
     <name>212 New York, NY, US</name>
   </entity>
  ▼<entity>
     <documentType>place</documentType>
     <woeid>2388929</woeid>
     <type>Town</type>
     <name>Dallas, TX, US</name>
   </entity>
  ▼<entity>
     <documentType>ancestor</documentType>
     <woeid>2347591</woeid>
     <type>State</type>
     <name>New York</name>
   </entity>
  ▼<entity>
     <documentType>administrativeScope</documentType>
     <woeid>2459115</woeid>
     <type>Town</type>
     <name>New York, NY, US</name>
   </entity>
 </geo>
 ▼<temp>
    <dateDoc>[??:02/05/2004:??.??:??]</dateDoc>
    <date>[X:??/??/??:??.??:??]</date>
    <date>[??:??/3/??:??.??:??]</date>
    <date>[??:??/3/??:??.??:??]</date>
    <date>[G:??/??/??:??.??:??]</date>
    <date>[G:??/??/??:??.??:??]</date>
 </temp>
 ▼<names>
    <name/>
    <name>Jenkens jenkens</name>
    <name>Gilchrist gilchrist</name>
    <name>IRS irs</name>
    <name>IRS irs</name>
    <name>Jenkens jenkens</name>
    <name>Gilchrist gilchrist</name>
    <name>New_York new_york</name>
    <name>David_Deary david_deary</name>
 </names>
</document>
```

Figura 2: Ejemplo de fichero *XML* de filtro creado a partir de un documento del corpus.

Geográfica: en este apartado se guarda la parte geográfica relevante del artículo original. Esta información geográfica es extraída mediante *Yahoo! Placemaker*. Concretamente se guardan los siguientes datos por cada uno de los topónimos localizados en el texto: tipo (indica si el topónimo que describe es el ámbito genérico del texto, una entidad administrativa superior, un lugar encontrado en el texto, etc.), *WOEID*, tipo de topónimo (indica si el lugar encontrado es un país, una ciudad, una entidad vaga, etc.), y nombre.

Temporal: en este grupo se guardan todas las fechas (en formato normalizado) encontradas en el texto original. Al menos tendrá una entrada, la fecha del artículo, y detrás de ésta vendrán el restos de fechas que se hayan localizado en el texto.

Entidades: esta sección es la que recoge todas las entidades no geográficas nombradas en el artículo original.

3.2.2. Consultas.

Se han creado las siguientes secciones por cada una de las consultas enviadas (figura 3):

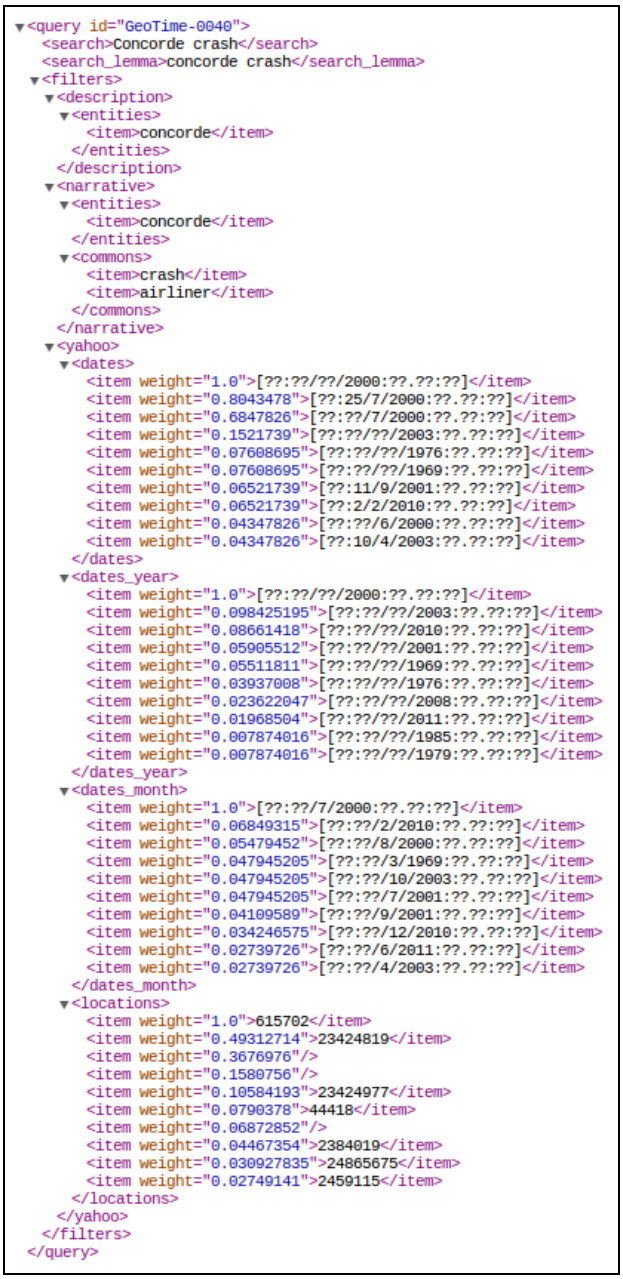
- Términos de búsqueda: Todos los términos de búsqueda, sin *stopwords*.
- Términos de búsqueda lematizados: lo mismo del apartado anterior pero lematizado.
- Filtros:
  - Parte descriptiva: fechas, topónimos y entidades encontradas en la parte descriptiva de la consulta.
  - Parte narrativa: análogamente, contendrá los mismos tres apartados descritos en la parte descriptiva más restricciones de topónimos y temporales.
  - Extensión de consulta: contendrá entradas expandidas de los términos más representativos de la consulta para una posible extensión de la misma.
  - Q&A: parte que contendrá los datos extraídos de la búsqueda de respuestas mediante *Yahoo!*, como fechas (completas o incompletas), fechas con mes y año, fechas con año, y topónimos. Tendrá los 10 valores más significativos normalizados por cada uno de los datos mencionados anteriormente.

3.3. Funcionamiento del sistema

El funcionamiento del sistema *GIR* se divide en tres fases: una inicial que se encarga de indexar y preprocesar todo el corpus, la segunda que procesa las consultas, y una final que es la encargada de ejecutar dichas consultas.

3.3.1. Preprocesado del corpus.

En esta fase se indexa el corpus lematizado con *Lucene* y *Terrier*, se obtienen las entidades geográficas con *Yahoo! Placemaker*, y se obtienen las entidades nombradas y temporales con



```
<query id="GeoTime-0040">
  <search>Concorde crash</search>
  <search_lemma>concorde crash</search_lemma>
  <filters>
    <description>
      <entities>
        <item>concorde</item>
      </entities>
    </description>
    <narrative>
      <entities>
        <item>concorde</item>
      </entities>
      <commons>
        <item>crash</item>
        <item>airliner</item>
      </commons>
    </narrative>
    <yahoo>
      <dates>
        <item weight="1.0">[??:??/??/2000:??.??:??]</item>
        <item weight="0.8043478">[??:25/7/2000:??.??:??]</item>
        <item weight="0.6847826">[??:??/7/2000:??.??:??]</item>
        <item weight="0.1521739">[??:??/??/2003:??.??:??]</item>
        <item weight="0.07608695">[??:??/??/1976:??.??:??]</item>
        <item weight="0.07608695">[??:??/??/1969:??.??:??]</item>
        <item weight="0.06521739">[??:11/9/2001:??.??:??]</item>
        <item weight="0.06521739">[??:2/2/2010:??.??:??]</item>
        <item weight="0.04347826">[??:??/6/2000:??.??:??]</item>
        <item weight="0.04347826">[??:10/4/2003:??.??:??]</item>
      </dates>
      <dates_year>
        <item weight="1.0">[??:??/??/2000:??.??:??]</item>
        <item weight="0.098425195">[??:??/??/2003:??.??:??]</item>
        <item weight="0.08661418">[??:??/??/2010:??.??:??]</item>
        <item weight="0.05905512">[??:??/??/2001:??.??:??]</item>
        <item weight="0.05511811">[??:??/??/1969:??.??:??]</item>
        <item weight="0.03937008">[??:??/??/1976:??.??:??]</item>
        <item weight="0.023622047">[??:??/??/2008:??.??:??]</item>
        <item weight="0.01968504">[??:??/??/2011:??.??:??]</item>
        <item weight="0.007874016">[??:??/??/1985:??.??:??]</item>
        <item weight="0.007874016">[??:??/??/1979:??.??:??]</item>
      </dates_year>
      <dates_month>
        <item weight="1.0">[??:??/7/2000:??.??:??]</item>
        <item weight="0.06849315">[??:??/2/2010:??.??:??]</item>
        <item weight="0.05479452">[??:??/8/2000:??.??:??]</item>
        <item weight="0.047945205">[??:??/3/1969:??.??:??]</item>
        <item weight="0.047945205">[??:??/10/2003:??.??:??]</item>
        <item weight="0.047945205">[??:??/7/2001:??.??:??]</item>
        <item weight="0.04109589">[??:??/9/2001:??.??:??]</item>
        <item weight="0.034246575">[??:??/12/2010:??.??:??]</item>
        <item weight="0.02739726">[??:??/6/2011:??.??:??]</item>
        <item weight="0.02739726">[??:??/4/2003:??.??:??]</item>
      </dates_month>
      <locations>
        <item weight="1.0">615702</item>
        <item weight="0.49312714">23424819</item>
        <item weight="0.3676976"/>
        <item weight="0.1580756"/>
        <item weight="0.10584193">23424977</item>
        <item weight="0.0790378">44418</item>
        <item weight="0.06872852"/>
        <item weight="0.04467354">2384019</item>
        <item weight="0.030927835">24865675</item>
        <item weight="0.02749141">2459115</item>
      </locations>
    </yahoo>
  </filters>
</query>
```


Figura 3: Ejemplo de fichero *XML* creado a partir de una consulta.

FreeLing. Con toda esta información se crea un fichero *XML* por cada documento del corpus, que se empleará a la hora de valorar la relevancia de los documentos con respecto a la consulta en la fase de ejecución (figura 2).

3.3.2. Procesado de las consultas.

En esta segunda fase, las consultas son enviadas al módulo de análisis lingüístico para obtener la información descrita anteriormente en dicho módulo. Una vez finalizado el trabajo en el módulo de análisis lingüístico, el sistema envía la consulta al módulo de Q&A. Seguidamente, con todos los resultados obtenidos de los dos módulos anteriores, el sistema envía cada una de las referencias geográficas encontradas al módulo geográfico, el cual transformará cada una de estas referencias en el identificador unívoco de *Yahoo! Placemaker* (*WOEID*). Por último, el sistema almacena todos estos datos creando un nuevo documento XML por cada una de las consultas leídas (figura 3).

3.3.3. Ejecución.

En esta tercera y última fase, el sistema obtiene los archivos *XML* de las consultas y, junto con los *XML* de los documentos relevantes recuperados se las envía al módulo evaluador que ejecutará la tarea descrita en dicho módulo. Una vez finalizada la reordenación de los documentos relevantes para cada una de las consultas, el módulo evaluador guardará los resultados en un fichero de soluciones.

4. Experimentos y evaluación

En esta sección se describirán por un lado las métricas utilizadas para la evaluación del sistema así como el entorno donde se realizó dicha evaluación, y por otro lado el impacto de cada uno de los componentes del sistema.

4.1. Métricas y entorno de evaluación

A continuación se puede ver dónde se ha evaluado el sistema descrito en este trabajo y se razonará la elección de una métrica de evaluación de los resultados.

4.1.1. Entorno de evaluación

Por un lado, en una primera fase, se ha evaluado el sistema aquí descrito con las colecciones de documentos y preguntas de la tarea *GeoTime* del *NTCIR 2010*. Esta tarea combina *GIR* con búsqueda basada en el tiempo para encontrar eventos específicos en una colección de documentos. El conjunto de evaluación lo forman un corpus de documentos (artículos del *New York Times 2002-2005*) y un corpus de 25 preguntas, las cuales incluyen una parte descriptiva (p. ej. "*When and where did Astrid Lindgren die?*") y una parte narrativa ("*The user wants to know when and in what city the children's author Astrid Lindgren died.*"). Al tratarse de un *workshop* oriental, los organizadores trataron de organizar las tareas haciendo referencia a cualquier tipo de lengua asiática, aceptando como única lengua occidental el inglés, ya que es la lengua común para este tipo de campo de investigación. Debido a que no dominábamos ninguna lengua oriental, se optó por trabajar únicamente en inglés.

En una segunda fase, se evaluó el sistema en la tarea *GeoTime* del *NTCIR 2011*. Para esta nueva edición del *NTCIR 2011 GeoTime* se emplearon otras 25 preguntas con las mismas características

que las de la edición del año anterior. En cuanto al corpus empleado, continuaron empleándose los artículos del *New York Times 2002-2005* y se añadieron artículos de 3 corpora más, *Mainichi Daily 1998-2001*, *Korea Times 1998-2001* y *Xinhua English 1998-2001* y, al igual que con la tarea del año anterior, se optó por trabajar únicamente con corpora y consultas en inglés.

4.1.2. Métricas de evaluación

Para el análisis de los resultados expuestos en esta memoria se ha optado por la utilización de la métrica *nDCG*[13] (*normalized Discounted Cumulative Gain*) (Järvelin y Kekäläinen, 2002) utilizada en el *NTCIR 2010* y en el *NTCIR 2011* para poder comparar los sistemas participantes entre sí. Se ha optado por esta métrica de entre las tres empleadas en el *NTCIR* (sección 2.7), ya que ésta es una de las que son capaces de hacer evaluaciones graduales, es decir, que no son simplemente decisiones binarias entre válido y no válido, como hace la métrica *AP*, por ejemplo, si no que es capaz de evaluar un documento parcialmente relevante, como se hizo en las dos ediciones del *workshop* mencionado. Estas evaluaciones podían ser: completamente relevante, relevante geográficamente, relevante temporalmente, relevante de algún modo, e irrelevante. En el *NTCIR-GeoTime* esta métrica fue utilizada en tres bases diferentes: 10, 100 y 1.000. Se ha escogido la base 1.000 en los experimentos aquí mostrados (la misma base que el número de documentos recuperados por consulta), lo que significa que no se está teniendo en cuenta la posición de los documentos recuperados, si no que se está considerando si los documentos recuperados son de algún modo relevantes o no (ganancia acumulativa). Esto se ha hecho debido a que nos estamos centrando más en obtener el mayor porcentaje posible de documentos relevantes, más que en el orden correcto, ya que esto último se realizará en un trabajo futuro donde se emplearán los módulos ya existentes para obtener una puntuación por documento más precisa de los documentos recuperados inicialmente por el módulo del motor de búsqueda.

4.2. Impacto de los componentes

Los experimentos fueron elaboradas según los esquemas de pesados descritos en el módulo de reordenación (sección 3.1.7 en página 7). Para ambos esquemas fueron asignando distintos valores a las variables λ, α y β, realizado un recorrido sistemático sobre estos valores en los intervalos [0,1] en incrementos de 0,1.

4.2.1. Motor de búsqueda

Como ya se ha comentado en la sección 4.1.1, en una primera fase de la experimentación las consultas y el corpus utilizados fueron los del *NTCIR 2010*. Para esta fase, el motor de búsqueda utilizado fue únicamente *Lucene*, obteniendo como resultado los mostrados en la figura 4 para los mejores valores de los parámetros α y β de los esquemas de pesado vistos en la sección 3.1.7.

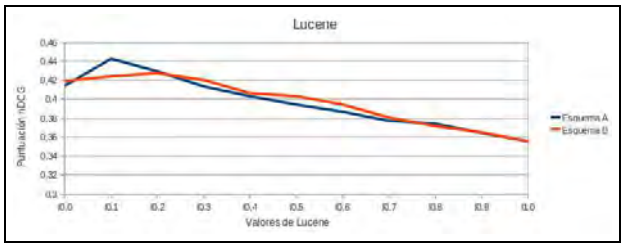


Figura 4: Gráfica de resultados de los esquemas de pesado A y B, desde la no utilización del ranking de *Lucene* (valor 0), hasta la utilización única y completa de éste (valor 1).

En la segunda fase de experimentación, la que tuvo lugar con los corpora del *NTCIR 2011*, se observó que la cobertura alcanzada por *Lucene* apenas superaba el 50 %, por lo que se llevo a cabo un experimento para comprobar qué hubiera sucedido si se hubiese alcanzado una mayor cobertura, cuyos resultados se pueden observar en la tabla 1. Estos resultados han sido clasificados en tres grupos:

1. Consultas que obtienen una cobertura entre el 0 % y el 100 %, es decir, todas las consultas.
2. Consultas que obtienen una cobertura entre el 50 % y el 100 % (12 consultas).
3. Consultas que obtienen una cobertura entre el 75 % y el 100 % (10 consultas).

En cada uno de estos tres grupos se pueden observar el tanto por cien de la cobertura de documentos relevantes recuperado por cada consulta, y la puntuación *nDCG-1000* alcanzada para la mencionada consulta. Finalmente, se obtiene la cobertura y puntuación media para las consultas que entran en cada uno de los tres grupos. Como ya se ha mencionado anteriormente, el objetivo de este experimento era comprobar lo que sucedería si se hubiese obtenido una mayor cobertura por parte del módulo del motor de búsqueda,

[13] *nDCG* mide la media normalizada de utilidad, o ganancia, de un documento basado en la posición en el ranking final de resultados. La ganancia es acumulativa desde lo más alto de la lista de resultados hasta el final, con la ganancia de cada resultado descontada al nivel que le sucede.

Tabla 1: Cobertura y puntuación *nDCG-1000* alcanzada utilizando únicamente *Lucene* por cada una de las consultas del *NTCIR 2011*.

	0 % - 100 %		50 % - 100 %		75 % - 100 %	
Topic	Cobertura	nDCG	Cobertura	nDCG	Cobertura	nDCG
GeoTime-0026	93,2945 %	0,7730	93,2945 %	0,7730	93,2945 %	0,7730
GeoTime-0027	85,7143 %	0,2576	85,7143 %	0,2576	85,7143 %	0,2576
GeoTime-0028	85,4839 %	0,5846	85,4839 %	0,5846	85,4839 %	0,5846
GeoTime-0029	43,3566 %	0,2806	-	-	-	-
GeoTime-0030	66,6667 %	0,3467	66,6667 %	0,3467	-	-
GeoTime-0031	36,6667 %	0,2905	-	-	-	-
GeoTime-0032	35,0877 %	0,3367	-	-	-	-
GeoTime-0033	74,4186 %	0,5660	74,4186 %	0,5660	-	-
GeoTime-0034	86,3636 %	0,4655	86,3636 %	0,4655	86,3636 %	0,4655
GeoTime-0035	28,5714 %	0,1031	-	-	-	-
GeoTime-0036	31,9149 %	0,2849	-	-	-	-
GeoTime-0037	0,0000 %	0,0000	-	-	-	-
GeoTime-0038	1,6908 %	0,0317	-	-	-	-
GeoTime-0039	84,1202 %	0,6174	84,1202 %	0,6174	84,1202 %	0,6174
GeoTime-0040	82,0755 %	0,7887	82,0755 %	0,7887	82,0755 %	0,7887
GeoTime-0041	98,9362 %	0,7117	98,9362 %	0,7117	98,9362 %	0,7117
GeoTime-0042	1,2739 %	0,0145	-	-	-	-
GeoTime-0043	91,4894 %	0,5294	91,4894 %	0,5294	91,4894 %	0,5294
GeoTime-0044	28,5714 %	0,1920	-	-	-	-
GeoTime-0045	75,0000 %	0,6110	75,0000 %	0,6110	75,0000 %	0,6110
GeoTime-0046	92,3077 %	0,7454	92,3077 %	0,7454	92,3077 %	0,7454
GeoTime-0047	6,6667 %	0,0174	-	-	-	-
GeoTime-0048	47,9167 %	0,4963	-	-	-	-
GeoTime-0049	60,0000 %	0,6509	60,0000 %	0,6509	-	-
GeoTime-0050	57,1429 %	0,2031	57,1429 %	0,2031	-	-
Cobertura media	**55,7892 %**		**80,9295 %**		**87,4785 %**	
Puntuación media	**0,3959**		**0,5607**		**0,6081**	

pudiéndose apreciar la sustancial mejora obtenida en las dos últimas filas de la tabla 1 (pasando de una puntuación de 0,3959 a 0,5607 o 0,6081, según la cobertura mínima requerida).

Observando dichos resultados, y basándonos en el trabajo realizado por (Perea-Ortega, 2010), se incorporó un motor de búsqueda adicional, *Terrier*. Mediante la utilización de ambos motores de búsqueda, se pasó del 55,7892 % de cobertura al 87,0165 %. Dicha cobertura hizo que la puntuación obtenida pasara de 0,3959 a 0,5921 (*nDCG-1000*) utilizando únicamente los motores de búsqueda.

4.2.2. Módulo de análisis lingüístico

En lo que al módulo de análisis lingüístico respecta, como ya se ha mencionado previamente en la sección 3.1.2, es el encargado de procesar sintácticamente el contenido de cada uno de los artículos del corpus, por lo que para evaluar su funcionamiento se optó por ver cuan importante era la parte descriptiva de las consultas respecto a la narrativa. Para ello se utilizó el esquema de pesado A visto en la ecuación 1, asignándole el

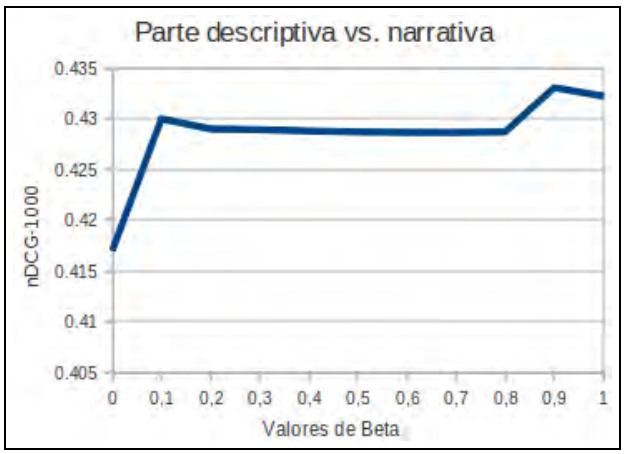


Figura 5: Importancia de la parte narrativa de la consulta contra la descriptiva sobre las consultas del *NTCIR 2011*.

valor que mejor resultados obtuvo a la variable α e incrementando gradualmente los valores de β desde 0 hasta 1, tal y como se puede apreciar en la figura 5. Esta gráfica nos muestra cómo es sustancialmente más relevante para nuestro sistema la parte descriptiva de la consulta frente a la narrativa, lo que conduce a pensar que este módulo debe ser mejorado para extraer así mejor las características más importantes descritas en la parte narrativa de la consulta, las cuales, ahora mismo no se tienen en cuenta.

4.2.3. Módulo de Q&A

Sobre los resultados del *NTCIR 2011*, y utilizando únicamente el motor de búsqueda *Lucene*, se realizó un experimento para evaluar la importancia del módulo de Q&A. Para ello se utilizó el esquema de pesado A (ecuación 1) explicado en la sección 3.1.7, asignándole el valor que mejores resultados obtiene a la variable β y variando los valores de α para comprobar dicha importancia. Cómo se puede ver en la figura 6, para el intervalo de valores que va de 0,2 a 0,9 los resultados se mantienen muy igualados, lo que parece decir que el módulo de Q&A tiene importancia aunque no sea crucial en el resultado final.

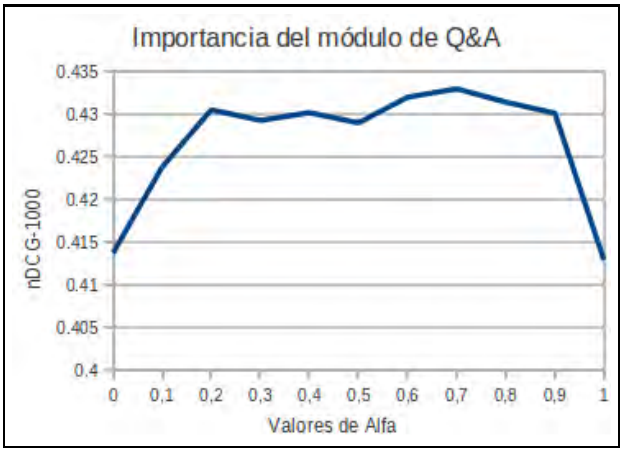


Figura 6: Importancia del módulo de Q&A sobre las consultas del *NTCIR 2011*.

Se realizó un estudio más exhaustivo sobre este módulo y se observó que los documentos *XML* creados tras el tratamiento de las consultas (figura 3), en el apartado que concierne a este módulo, en la inmensa mayoría de las ocasiones se contestaba a la parte temporal y/o geográfica de la consulta, por lo que se decidió llevar a cabo un experimento donde se expandiera la consulta lanzada al motor de búsqueda *Lucene* los 10 términos del apartado de fechas, completas o incompletas (*dates*), y los 10 términos del apartado de topónimos (*locations*), todos ellos con sus respectivos pesos. Posteriormente, se unirían los documentos devueltos por *Lucene* con los devueltos por *Terrier*, tal y cómo se explicó en la sección 3.1.1. Como resultado de este experimento se pasó de una puntuación *nDCG-1000* de 0,5921 a 0,6206.

4.2.4. Módulo de análisis temporal y módulo geográfico

Se ha comprobado el peso que tiene el módulo de análisis temporal en el sistema frente al módulo geográfico. Dicha comprobación se ha realizado en el marco de las consultas del *NTCIR 2011*, utilizando para ello el esquema de pesado B (sección 3.1.7) reflejado en la ecuación 2. Para este análisis, se le ha asignado el mejor valor obtenido para la variable α y se ha ido incrementando gradualmente el valor de la variable β en el intervalo que va desde 0 hasta 1, correspondiendo el valor 0 a la utilización exclusiva de la parte temporal frente a la geográfica y el valor uno lo contrario. La figura 7 muestra los resultados, donde se puede observar cómo la parte geográfica adquiere una mayor importancia que la temporal.

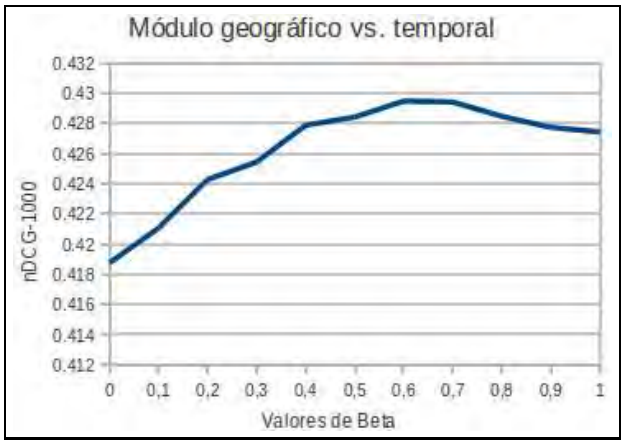


Figura 7: Importancia de los módulo geográfico y temporal sobre las consultas del *NTCIR 2011*.

4.2.5. Módulo para la detección de entidades

Uno de los análisis más interesantes que cabe hacer de los resultados es el del comportamiento de las consultas individualmente para la mejor configuración del sistema.

Dentro del marco del *NTCIR 2010*, en la figura 8, se puede apreciar como hay algunas consultas para las que el sistema obtuvo un comportamiento excepcional, mejorando claramente al mejor de los sistemas de los que participaron en el *GeoTime 2010*. Estos resultados suelen darse en consultas que tienen alguna entidad (nombres de persona, de organizaciones, etc.) poco común, es decir, que aparecen rara vez en el corpus dado. De igual manera, hay algunas consultas para las cuales el sistema obtiene un pobre resultado. Estos acostumbran a darse cuando se encuentran

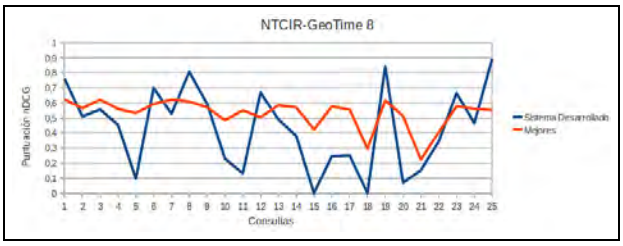


Figura 8: Gráfica de resultados por cada una de las 25 consultas del *NTCIR 2010*, comparando el sistemas expuesto con los mejores resultados para estas consultas de entre todos los participantes.

términos muy comunes en las consultas.

Debido a lo anteriormente expuesto, se decidió obtener la gráfica que mostrara la importancia que tiene el módulo de detección de entidades sobre las consultas del *NTCIR 2011*, cuyo resultado se puede observar en la figura 9. Dicha gráfica ha sido obtenida mediante la utilización del esquema de pesado *B*, plasmado en la ecuación 2, asignándole su mejor valor a la variable α, y obteniendo valores para la variable β dentro del intervalo 0,1. La gráfica nos muestra la gran importancia que tiene el trato de las entidades (excluyendo las geográficas y las temporales) en el corpus. Se puede apreciar como a medida que dejamos de utilizar exclusivamente las entidades detectadas (valor $\beta = 0$) para ir utilizando únicamente las entidades geográficas y temporales (valor $\beta = 1$), los resultados empeoran pese a ser unas consultas geo-temporales.

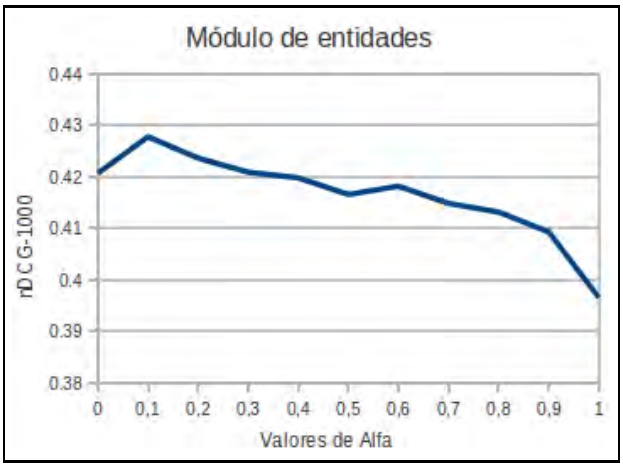


Figura 9: Importancia del módulo de detección de entidades sobre las consultas del *NTCIR 2011*.

4.2.6. Módulo de reordenación

La importancia de este módulo puede verse reflejada en la figura 4, donde se puede ver cómo al reordenar los resultados para las consultas del *NTCIR 2010* devueltas por *Lucene*, mejoran los resultados.

Por otro lado, al introducir las técnicas de Q&A descritas en el experimento llevado a cabo en la sección 4.2.3, y después de añadir el motor de búsqueda *Terrier*, tal y como se pudo ver en la experimentación realizada en 4.2.1, la reordenación ha sido implementada realizando una intersección de los resultados devueltos por ambos motores de búsqueda, tal y como se se explicó en la sección 3.1.1, por lo que prácticamente la reordenación la llevan a cabo dichos motores. Esta reordenación es crucial para cualquier sistema de *IR*, por lo que no se puede plantear la existencia de un sistema *GIR* sin la existencia de ésta.

5. *Conclusiones y trabajo futuro*

En este trabajo se ha realizado una introducción al campo de la recuperación de información geográfica, implementando un sistema *GIR* propio y evaluándolo en el contexto de una competición internacional como es el *NTCIR*.

En primer lugar, cabe destacar la gran mejora que realiza el sistema con todos sus módulos con respecto a un simple sistema de *IR*, alrededor de 10 puntos porcentuales (figura 4), por lo que, para consultas que contengan un perfil geográfico, dicho sistema supone un gran aumento en la relevancia de los resultados. Dicha mejora se ha conseguido utilizando prácticamente sólo la parte descriptiva de las consultas, ya que la parte narrativa de las mismas introducía ruido en abundancia. En un futuro habría que mejorar el módulo de análisis lingüístico para que sea capaz de extraer y/o filtrar mejor la información de la parte narrativa de las consultas.

Si nos centramos en los módulos del sistema, cabe destacar el gran funcionamiento que ha tenido el módulo de Q&A. Observando esto, es obligatorio resaltar que como uno de los posibles trabajos futuros habría que investigar técnicas más avanzadas en Q&A.

Por otro lado, como se ha podido ver en la sección 4.2.4, el módulo temporal ha tenido menor peso que el geográfico. Debido a esto se está sopesando la incorporación al sistema de un módulo temporal más completo, para lo cual se está barajando la opción del sistema *TIPSem*[14] desarrollado en el *GPLSI* de la *Universidad de Alicante*.

Centrándonos en el módulo puramente geográfico, actualmente se tienen dos frentes abiertos. Por un lado, la obtención de más metadatos de *Yahoo! Placemaker*, como pueden ser el ámbito general del que habla el documento analizado (la huella o *footprint*) para su posterior tratamiento.

El objetivo final sería el implementar un siste-

[14] `http://gplsi.dlsi.ua.es/demos/TIMEE/`

ma completo que satisfaga todos los puntos descritos en la sección 2.

6. Agradecimientos

Esta investigación ha sido parcialmente financiada por el gobierno de España bajo del proyecto TEXTMESS 2.0 (TIN2009-13391-C04-01), y por la Universidad de Alicante bajo el proyecto GRE10-33.

Bibliografía

Baeza-Yates, Ricardo y Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*, volumen 463. Addison Wesley.

Cardoso, Nuno y Diana Santos. 2007. To separate or not to separate : reflections about current gir practice. *English*.

Clough, Paul. 2005. Extracting metadata for spatially-aware information retrieval on the internet. En Chris Jones y Ross Purves, editores, *GIR*, páginas 25–30. ACM.

Järvelin, Kalervo y Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20:422–446, October.

Jones, C B, R S Purves, P D Clough, y H Joho. 2008. Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science*, 22(10):1045–1065.

Jones, Christopher B y Ross S Purves. 2008. Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228.

Jones, Christopher Bernard, A Arampatzis, P Clough, y R S Purves. 2007. The design and implementation of spirit: a spatially-aware search engine for information retrieval on the internet.

Leveling, J y S Hartrumpf. 2007. On metonymy recogntion for geographic ir. En *Proceedings of GIR2006 the 3rd Workshop on Geographical Information Retrieval*.

Li, Yi, Alistair Moffat, Nicola Stokes, y Lawrence Cavedon. 2006. Exploring probabilistic toponym resolution for geographical information retrieval. En Ross Purves y Chris Jones, editores, *GIR*. Department of Geography, University of Zurich.

Mitamura, Teruko, Eric Nyberg, Hideki Shima, Tsuneaki Kato, Tatsunori Mori, Chin yew Lin, Ruihua Song, Chuan jie Lin, Tetsuya Sakai, y Donghong Ji Noriko K. 2008. Overview of the ntcir-7 aclia tasks: Advanced cross-lingual information access.

Montello, D R, M F Goodchild, Jonathon Gottsegen, y Peter Fohl. 2003. Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition Computation*, 3(2):185–204.

Perea-Ortega, J.M. 2010. Recuperación de información geográfica basada en múltiples formulaciones y motores de búsqueda. *Procesamiento del Lenguaje Natural. N. 46 (2011). ISSN 1135-5948*, páginas 131–132.

Robertson, Stephen E, Steve Walker, y Micheline Hancock-Beaulieu, 1998. *Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive*, páginas 199–210. NIST.

Silva, Mário J, Bruno Martins, Marcirio Chaves, Ana Paula Afonso, y N Cardoso. 2006. Adding geographic scopes to web resources. *Computers Environment and Urban Systems*, 30(4):378–399.

Vaid, S, Christopher Bernard Jones, H Joho, y M Sanderson. 2005. Spatio-textual indexing for geographical search on the web. *Advances in Spatial and Temporal Databases 9th International Symposium SSTD 2005*, 3633:218–235.

Van Kreveld, Marc, Iris Reinbacher, Avi Arampatzis, y Roelof Van Zwol. 2005. Multidimensional scattered ranking methods for geographic information retrieval*. *Geoinformatica*, 9:61–84, March.

Wang, Chuang, Xing Xie, Lee Wang, Yansheng Lu, y Wei-Ying Ma. 2005. Detecting geographic locations from web resources. En Chris Jones y Ross Purves, editores, *GIR*. ACM.

Zhang, Vivian Wei, Benjamin Rey, y Rosie Jones. 2006. Geomodification in query rewriting. En *In Proceeding of the 3 rd workshop of Geographic Information Retrieval*.