# Ordenamiento Basado en Ejemplos para la Recuperación de Información Geográfica

por

**Esaú Villatoro Tello**
MC., INAOE

Tesis sometida como requisito parcial para obtener el grado de

**DOCTOR EN CIENCIAS EN LA ESPECIALIDAD DE CIENCIAS COMPUTACIONALES**

en el

**Instituto Nacional de Astrofísica, Óptica y Electrónica**
Octubre 2010
Tonantzintla, Puebla

Supervisada por:

**Dr. Luis Villaseñor Pineda**
Investigador Titular del INAOE

**Dr. Manuel Montes y Gómez**
Investigador Titular del INAOE



# Resumen

Las técnicas de recuperación de información actuales representan un avance importante para tratar con el problema del exceso de información. Los motores de búsqueda en la Web son ejemplos convincentes de su utilidad y popularidad. Gracias a esto, para los usuarios de la Web, la tarea de recuperación de información se ha vuelto una actividad cotidiana. Los motores de búsqueda son utilizados para localizar información referente a casi todos los dominios del conocimiento humano. Ahora bien, gran parte de toda esta información está acotada o es considerada como perteneciente a un espacio geográfico, es decir, en su interior, se hace referencia a diferentes aspectos espaciales y/o datos geográficos, como consecuencia muchos usuarios especifican una referencia geográfica (nombre de una ciudad, provincia, avenida, etc.) como parte de su consulta al buscador.

Dentro de esta tesis nos enfocaremos en el manejo de este tipo de consultas, por ejemplo, usuarios buscando departamentos en renta, reservación de hoteles, restaurantes, información sobre sitios arqueológicos, atractivos turísticos, sitios industriales, universidades, etc. Problema que intenta resolver un área conocida como Recuperación de Información Geográfica (GIR).

Avances recientes en el área de recuperación de información geográfica han mostrado que el problema está parcialmente resuelto a través de técnicas tradicionales de recuperación de información (IR). Se ha podido observar que máquinas tradicionales de IR son capaces de recuperar la mayoría de los documentos relevantes para la mayor parte de las consultas geográficas, sin embargo, presentan dificultades al momento de generar un orden pertinente de los documentos recuperados, lo cual resulta en un desempeño deficiente. Una de las razones por las que el ordenamiento es inapropiado es la falta de información en la consulta. Es por esta razón que grupos de investigación

han tratado de cubrir esta falta de información empleando recursos geográficos robustos (e.g. ontologías geográficas), mientras que algunos otros grupos de investigación lo han intentado por medio de estrategias de reformulación de consultas a través de técnicas de retroalimentación de relevancia.

Como una alternativa, en este trabajo proponemos una estrategia re-ordenamiento para sistemas GIR. Dado que máquinas de IR son capaces de recuperar un porcentaje considerable de documentos relevantes a consultas geográficas, nuestro trabajo se enfoca en el proceso de mejorar el orden asignado inicialmente a un conjunto de documentos recuperados por medio de emplear información obtenida a través de un proceso de retroalimentación, es decir, se propone *un re-ordenamiento vía retroalimentación de relevancia*. Adicionalmente, dado que las consultas geográficas tienden a incluir información implícita, nosotros proponemos el uso de documentos completos en lugar de términos aislados en el proceos del re-ordenamiento, a los cuales denominamos *documentos ejemplo*.

El objetivo de la estrategia propuesta es aproximar a una forma más explícita los requerimientos de información implícitos contenidos en las consultas geográficas, y como consecuencia otorgar un orden más pertinente al conjunto de documentos recuperados. Los resultados experimentales muestran que el método propuesto logra ser mejor al momento de generar un orden de los documentos, superando al método base en un 5.4 % en el mejor de los casos bajo un esquema de retroalimentación ciega, mientras que para el caso de retroalimentación simulada se logró tener una ventaja de hasta un 27 % con respecto al método base.

# Abstract

Current information retrieval techniques represent an important improvement for the problem of dealing with the excess of information. Web's search machines are convincing examples of their utility and popularity. Hence, Web's users see the task of information retrieval as an every day activity. Search machines are employed to find information about almost every area of knowledge. However, a considerable part of this information is known to be geographically delimited, that is, in their content refers to some spatial aspects, or some geographical places (for instance, names of cities, countries, towns or streets, etc.). As a consequence of this fact, users looking for this type of information tend to include in their queries some geographical references when they are doing some search in the Web.

In this work we focus on this particular type of queries, this is, users looking for apartments, hotels, restaurants, archeological information, touristic activities, industrial places, interchange universities, etc. This problem is known as Geographic Information Retrieval.

Recent research in the area of Geographic Information Retrieval (GIR) has shown that the problem is partially solved through traditional Information Retrieval (IR) techniques. It has been possible to observe that traditional IR machines are able to retrieve the majority of relevant documents to some geographical queries; however, these IR machines are not capable of generating a pertinent ranking of retrieved documents, which turns out into a bad system performance. One of the main reasons for the inappropriate ranking is the lack of information in the given query. For this reason many research groups have tried to fill this lack of information employing robust geographical resources (e.g., geographic ontologies), while some others have tried to do the same by means of query reformulation strategies via relevance feedback.

As an alternative, in this work we propose a strategy for re-ranking the output of GIR systems. Given that retrieving relevant documents to geographic queries seems

to be a minor problem for traditional IR machines, our work focuses on the problem of improving the initial ranking of a set of documents by employing information obtained through a relevance feedback process, i.e., *re-ranking via relevance feedback*. Additionally, since geographic queries tend to include implicit information requirements, we propose the use of complete documents instead of isolated terms to perform the process of re-ranking; such documents are called *example documents*.

The aim of the proposed solution is to obtain an explicit approximated version of the implicit information requirements contained in the original geographic query, and therefore, to produce a more pertinent ranking of retrieved documents. Experimental results show that the proposed method is able to produce a more pertinent order among the retrieved documents. In the one hand, under a blind feedback scheme our method outperforms the baseline by a difference of 5.4 %; in the other hand, considering a simulated feedback scheme our method reaches an advantage of 27 % over the baseline.

# Índice General

# Lista de Figuras

# Lista de Tablas

# Capítulo 1
# Introducción

En las últimas décadas se ha presentado un crecimiento exponencial en la cantidad de textos disponibles en formato digital. Un claro ejemplo de este gran aumento de información es la Web, la cual día a día sigue aumentando en contenido así como en demanda. Este avance en tamaño ha originado el problema de cómo localizar, acceder y tratar toda esta inmensa cantidad de información.

Las técnicas de recuperación de información actuales representan un avance importante para tratar con el problema del exceso de información. Su capacidad para localizar documentos relevantes a una petición es sobresaliente. Los motores de búsqueda en la Web son ejemplos convincentes de su utilidad y popularidad.

Para los usuarios de la Web la tarea de recuperación de información se ha vuelto una actividad de todos los días. Los motores de búsqueda son utilizados para localizar información referente a casi todos los dominios del conocimiento humano. Sin embargo, gran parte de toda esta información puede ser considerada como perteneciente a un espacio geográfico (i.e., en su interior se hace referencia a diferentes aspectos espaciales y/o datos geográficos [77, 78, 93]), en consecuencia, muchos usuarios especifican una referencia geográfica (nombre de una ciudad, provincia, avenida, etc.) como parte de su consulta al buscador.

Este comportamiento se debe en gran parte a un fenómeno que sucede cada vez con mayor frecuencia, que es la necesidad de los usuarios por información precisa. Por ejemplo, supongamos que un usuario introduce una consulta como: "Turismo en el Noreste de Brasil". Tras esto, el usuario espera obtener documentos que hablen efectivamente de atractivos turísticos, y lo más importante es que sean atractivos turísticos dentro de la región Noreste de Brasil. Idealmente, los sistemas de recuperación de información deberán entregar un conjunto de documentos que estén semánticamente relacionados con la consulta hecha por el usuario, lo cual implicará satisfacer su ne-

cesidad de información.

Lo anterior contrasta con lo que los sistemas actuales de recuperación de información son capaces de hacer. Un sistema de recuperación de información tradicional entregará documentos que contengan únicamente las palabras de la consulta, por ejemplo, documentos que contengan la palabra "Turismo", "Noreste", "Brasil", y combinaciones de éstas. Sin embargo, la máquina de búsqueda no está considerando si la consulta se refiere a "Brasil" como país, o si se refiere al equipo de fútbol, o al nombre de alguna calle o avenida en un lugar distinto a Brasil. Y por si esto fuera poco, no se asegura que la relación espacial implícita en la consulta dada (i.e. turismo *dentro del Noreste de Brasil*) se mantenga en los documentos recuperados. En pocas palabras, muy probablemente la máquina de búsqueda entregará efectivamente documentos que hablen de "turismo" pero no todos necesariamente de "turismo en Brasil" y mucho menos de "turismo en la región Noreste de Brasil", lo que implica no satisfacer la necesidad de información del usuario.

En años recientes, diferentes grupos de investigación han enfocado sus esfuerzos en tratar de resolver este tipo de problema, conocido como *Recuperación de Información Geográfica*. En el Inglés, el idioma de interés dentro de esta propuesta, la precisión promedio de los sistemas de recuperación de información geográfica evaluados dentro del CLEF [1] fue de 26.31 %, 27.25 %, 25.55 % y 23.70 % en los años 2005, 2006, 2007 y 2008 respectivamente. Los resultados muestran que el avance ha sido prácticamente nulo. De aquí la necesidad de nuevas alternativas para abordar el problema.

## 1.1. Antecedentes

En la actualidad, mucha de la investigación que se ha realizado sobre métodos eficientes de acceso a información geográfica se ha enfocado únicamente en generar una representación geométrica del espacio (i.e., mapas digitales), basándose en el manejo de coordenadas [46]. Sin embargo, la mayoría de los usuarios cuando queremos hacer referencia a algún lugar geográfico lo hacemos empleando el nombre del lugar, y no por sus coordenadas específicas, lo cual es completamente natural pues el usuario común no piensa en esos términos.

Es conveniente mencionar que el énfasis que se le dió al uso de coordenadas es completamente entendible y justificado por la necesidad de recuperar, analizar y desplegar gráficamente información requerida por usuarios especializados en forma de

mapas digitales, con coordenadas y distintos datos espaciales específicos del lugar en cuestión. Sin embargo, conforme más fácil es el acceso a la información disponible en Internet, tanto para usuarios especializados como no-especializados, surge la creciente necesidad de contar con un sistema de recuperación de información que pueda ser capaz de manejar los conceptos de espacio en lenguaje natural [3].

Como se mencionaba anteriormente, mucha información disponible en la Web está geo-referenciada, es decir, es información identificada por medio de nombres de lugares u algunos otros datos geográficos (e.g., coordenadas geográficas). Esto ha permitido que sistemas actuales de recuperación de información generen índices basados en términos geográficos explícitos y/o en coordenadas geográficas con la intención de tratar de resolver las consultas geográficas. Sin embargo, existe información que no es directamente accesible por estos métodos, por ejemplo, supongamos la consulta: *"Malaria en los trópicos"*, donde el *nombre del lugar* no es explícito.

Para éstas y muchas de las diferentes consultas que un usuario proporcione no es necesario el producir un mapa digital o el utilizar información espacial detallada para satisfacer sus necesidades de información. En esta propuesta, nos enfocaremos en el manejo de este tipo de consultas, por ejemplo, usuarios buscando departamentos en renta, reservación de hoteles, restaurantes, información sobre sitios arqueológicos, atractivos turísticos, sitios industriales, universidades, etc.

Las máquinas de búsqueda que conocemos actualmente, realizan su tarea por medio de hacer un "emparejamiento" de palabras, es decir, entregan como documentos relevantes aquellos que contienen las palabras de la consulta inicial. Este enfoque es insuficiente para los propósitos de un sistema de recuperación de información geográfica, pues surge la necesidad de hacer una especie de "emparejamiento impreciso" entre los nombres de los lugares geográficos. Por ejemplo, si un usuario establece dentro de su consulta el nombre *Estados Unidos*, es necesario que el sistema de recuperación de información sepa identificar y relacionar a *Estados Unidos* con otros nombres con los cuales suele ser también referenciado este término, por ejemplo *EE.UU.*, *Estados Unidos de América*, *Norte América*, etc., incluso a regiones o ciudades dentro de él.

Dado lo anterior, podríamos pensar que si el sistema de recuperación de información geográfica tiene la capacidad adecuada para hacer este *emparejamiento impreciso* entre términos geográficos, obtendrá entonces una mayor cantidad de documentos posiblemente relacionados con la consulta inicial. En el ejemplo anterior se muestra de manera sencilla el problema que involucra el trabajar con nombres geográficos, sin

embargo, como se puede ver en la siguiente tabla (Tabla 1.1) hecha por Hauff et. al. [39], la diversidad de las consultas geográficas alcanza varios niveles.

| Nivel | Características | Ejemplo |
|---|---|---|
| **1** | Contienen el nombre concreto del lugar del que se desea información | *Comercio de diamantes en Angola* |
| **2** | Contienen el nombre de algún lugar y además con reglas simples sobre el lugar de relevancia | *Ciudades a no más de 100km alrededor de Frankfurt* |
| **3** | Contienen el nombre de algún lugar y además con reglas complejas sobre el lugar de relevancia | *Industria Automotriz alrededor del mar de Japón* |
| **4** | Contienen referencias muy generales de algún lugar, el cual no necesariamente aparece en un gazetter[1] | *Tropas Rusas en el sur del Cáucaso* |
| **5** | Contienen sólo referencias políticas del lugar de interés | *Créditos del antiguo Bloque del Este* |
| **6** | Contienen una descripción de características del lugar de interés | *Ciudades cercanas a volcanes activos* |

**Tabla 1.1: Clasificación de consultas geográficas por su complejidad**

En conclusión, para poder construir un sistema con las capacidades suficientes para resolver el problema de recuperación de información geográfica existen varios problemas interesantes a resolver. Algunos ejemplos de estos son:

- Arquitecturas para máquinas de búsqueda geográficas.
- Indexado espacial de documentos.
- Extracción de contextos geográficos de documentos y bases de datos geográficas.
- Técnicas de anotación geográfica para medios geo-referenciados.
- Diseño, construcción, mantenimiento y métodos de acceso a ontologías geográficas, gazetters[1] y tesauros geográficos.
- Interfaces para consultas geográficas en la Web.
- Visualización de resultados de búsquedas geográficas.
- Ordenamiento[2] por relevancia en búsquedas geográficas.

[1]Un gazetter es equivalente a hablar de un catálogo. En este caso hablamos de un catálogo de nombres geográficos.

[2]El *ordenamiento* u *ordenamiento por relevancia*, dentro del área de recuperación de información es también conocido como: *ranking* o *relevance ranking*

De manera particular, en este trabajo de investigación se propone atacar el problema del *ordenamiento por relevancia*, que como se verá en el capítulo 3, investigación reciente da evidencias de que resolviendo este problema máquinas de IR tradicionales podrían verse beneficiadas al momento de hacer búsquedas geográficas, además de que es uno de los problemas poco atacados.

## 1.2. Motivación

El adecuado manejo de información geográfica es la clave que permite a usuarios el realizar una planeación y toma de decisiones efectiva. Ejemplos de aplicaciones que ayudan en este proceso a los usuarios son los sistemas de posicionamiento global (GPS[3]), los cuales indican en todo momento la ubicación de un determinado objeto o persona. Otros ejemplos, son los sistemas de información geográfica (GIS[4]), o también conocidas como bases de datos geográficos (GDB[5]), los cuales permiten a un usuario acceder a un conjunto muy variado de información sobre un determinado lugar. Por ejemplo, si un usuario introduce el término "España" a un sistema GIS, éste le devolverá la ubicación geográfica de dicho término, generalmente en forma de un mapa digital junto con datos como: el número de habitantes, continente en el que se encuentra, dimensiones, código de país, países con los que hace frontera, etc[6].

Sin embargo, sin la necesidad de pensar en aplicaciones tan especializadas, podemos encontrar los famosos servicios Web (e.g, máquinas de búsqueda), los cuales permiten a todo tipo de usuarios (especialistas y no especialistas), tener acceso y además explorar gran cantidad de información. Este tipo de servicios tienen la particularidad de permitir a los usuarios satisfacer sus necesidades de información sin la necesidad de utilizar herramientas tan complejas y/o especializadas, razón por la que utilizar máquinas de búsqueda es algo tan común para un usuario de Internet. Esto es un comportamiento lógico, pues resulta más sencillo y natural para un usuario escribir una consulta del tipo *"Información sobre Madrid"* a poner algo como *"Latitud 40º 23' N, Longitud 3º 42' W"*[7], que sería un tipo de entrada que un sistema GIS o

[3]Global Positioning System
[4]Geographical Information System
[5]Geographical Data Base
[6]La cantidad de detalles incluidos en la información devuelta por un sistema GIS variará dependiendo del proveedor utilizado
[7]Coordenadas GPS de la ciudad de Madrid, España.

GPS esperaría.

Nótese que en el ejemplo anterior sólo se habla de *información sobre Madrid*, pero si se modifica la consulta a *"Coches bomba en Madrid"*, que sigue siendo un tipo de información, cualquier tipo de datos que un sistema GIS o GPS devuelva será totalmente irrelevante a las necesidades de información del usuario. El mismo problema ocurre con las máquinas de búsqueda actuales, pues al hacer consultas más complejas (e.g., Tipo 6, véase Tabla 1.1), que denotan una necesidad de información más específica; la precisión de los buscadores empieza a desmejorarse.

Por otra parte, nos enfrentamos a las preguntas: *¿Es realmente la recuperación de información geográfica un problema actual?*, i.e., *¿Hay usuarios que realmente hacen este tipo de consultas?, ¿Qué es lo que buscan?, ¿Con qué frecuencia las hacen?, ¿Qué características tienen este tipo de consultas?*.

En [77, 78, 93] se hacen diferentes estudios sobre conjuntos de consultas recopilados de diferentes buscadores en Internet. En estos estudios el objetivo principal fue analizar cómo formulan los usuarios sus consultas, qué es lo que buscan cuando introducen términos geográficos y de qué maneras describen una localidad geográfica.

Antes de pasar a describir los resultados obtenidos por estos trabajos, es conveniente definir qué es una consulta geográfica. Una consulta, puede ser considerada de tipo geográfica si contiene al menos uno de los siguientes elementos [78]:

- Nombres de lugares, e.g., Huston, Texas, US, etc.
- Algún tipo de información de localización, e.g., Códigos Postales.
- Adjetivos de lugar, e.g., americano, internacional, occidental
- Términos descriptivos del lugar, e.g, estado, ciudad, país, calle, etc.
- Atributos geográficos, e.g., lagos, islas, etc.
- Atributo de dirección, e.g., Norte, Sur, etc.

De manera más formal se dice que una consulta geográfica está formada por la tupla <*qué, relación, dónde*> [5]. Donde el *qué* es el término usado para especificar los términos generales (no-geográficos) de la necesidad de información, también llamada parte *temática* de la consulta. El *dónde* es el término utilizado para especificar las áreas geográficas de interés. Y la *relación* es el término que especifica la relación

espacial que conecta el *qué* con el *dónde*. Consultas complejas, que contienen múltiples relaciones espaciales, pueden ser vistas como combinaciones de estas tuplas[8].

En el estudio realizado por Sanderson y Kohler [78], se analizaron un total de 1,025,910 consultas proporcionadas por el buscador Excite[9]. Para el estudio tomaron de forma aleatoria una muestra de 2,500 consultas, de las cuales el 18 % contiene alguno de los atributos geográficos listados anteriormente. Además se hace un estudio sobre los tipos de *relaciones espaciales* más utilizadas. La palabra *"in"* es de las más utilizadas, seguida de *"at"* y *"from"*. El estudio muestra que los atributos de dirección *(north, south east, west)*, aunque con menor frecuencia, son utilizados para referirse a zonas geográficas grandes, e.g., países o estados. Por último, las relaciones *"near"* y *"surrounding"* son generalmente utilizadas cuando los usuarios buscan algo cerca de algún punto geográfico muy específico, i.e., búsquedas a nivel local.

En el trabajo realizado por Zhang et. al. [93] se analizaron un total de 4 millones de consultas, las cuáles fueron obtenidas del buscador de Yahoo!. En este trabajo, sólo se concentraron en identificar aquellas consultas que contenían el nombre de un lugar de manera explícita. Los resultados mostraron que cerca del 13 % de las consultas contenían en efecto el nombre de algún lugar. De manera adicional, el estudio mostró que en el 84 % de las veces los usuarios utilizan el nombre de ciudades, mientras que en un 13 % se hace uso del nombre de países y en un 3 % el nombre de estados y/o provincias.

Cabe mencionar que en este mismo trabajo se hizo un análisis sobre la forma en que los usuarios hacen la reformulación de sus consultas. El estudio mostró que los usuarios siempre modifican la parte geográfica de la consulta, dejando prácticamente intacta la parte temática. Este fenómeno se da en gran parte debido a la incapacidad del buscador de resolver la ambigüedad de los términos geográficos. Mostrando con esto que arquitecturas tradicionales no son suficientes para resolver este problema.

Por último, en el trabajo realizado por Sanderson y Han [77] se estudiaron 1 millón de consultas geográficas. El objetivo de este trabajo fue determinar con qué frecuencia los usuarios hacen uso de diferentes términos geográficos (nombre de países, ciudades, provincias, etc.). El estudio reveló que un 35 % de las veces los usuarios utilizan el nombre de un país, el de una región geográfica, estados y provincias; en un 21 % utilizan nombres de ciudades, y con menores porcentajes utilizan nombres compuestos

[8]Ejemplos de los niveles de complejidad de consultas geográficas se muestran en la Tabla 1.1
[9]http://search.excite.com/

de países o regiones.

En conclusión, estos trabajos muestran a diferentes niveles que las consultas geográficas son un sub-conjunto importante del total de consultas introducidas a un buscador en la vida real. También se muestra que, aunque se dan en un porcentaje importante este tipo de consultas, los usuarios casi siempre terminan reformulando la parte geográfica de sus consultas debido a que los buscadores no son capaces de satisfacer sus necesidades de información.

De aquí, el interés por atacar el problema de recuperación de información geográfica, dado que el proponer métodos efectivos, los cuales sean capaces de manejar los diferentes conceptos que se puedan encontrar dentro de una consulta geográfica, permitirá satisfacer las necesidades de información (cada vez más específicas) de los usuarios.

## 1.3. Descripción del Problema

Para los sistemas de Recuperación de Información la palabra *relevante* significa que los documentos recuperados deben estar semánticamente relacionados con la necesidad de información del usuario. Así entonces, uno de los principales problemas de IR es determinar qué documentos son, y cuáles no son documentos relevantes a una consulta. Posterior a la recuperación viene el problema del *ordenamiento por relevancia*, el cual tiene como objetivo principal el definir un orden entre los documentos recuperados de tal forma que los documentos más cercanos a la necesidad de información del usuario aparezcan en las primeras posiciones.

Los modelos de Recuperación de Información (IR) tales como el booleano, vectorial, probabilístico y modelos de lenguaje han representado a los documentos con un subconjunto de palabras clave (i.e., palabras de un índice) y han definido una función de *recuperación* para asociar con un grado de *relevancia* a cada documento con su respectiva consulta [7, 36]. En general, estos modelos han mostrado ser efectivos sobre varias tareas en diferentes foros de evaluación[10]. Sin embargo, la habilidad de estos modelos para ordenar documentos relevantes se ve limitada por la habilidad de los usuarios al momento de componer una consulta apropiada.

Con respecto a este hecho, modelos de IR tienden a fallar cuando los resultados deseados tienen requerimientos de información implícita que no son especificados en

[10]CLEF (http://www.clef-campaign.org/) y TREC (http://trec.nist.gov/)

las palabras clave de la consulta. Tal es el caso de la Recuperación de Información Geográfica (GIR), donde la búsqueda de documentos se basa no sólo en palabras clave temáticas, sino que también en palabras clave geográficas (e.g., palabras que hacen referencia a un lugar en el mundo) [48]. Por ejemplo, para la consulta "Ciudades cercanas a volcanes activos", los documentos esperados deberían mencionar nombres explícitos de ciudades y volcanes. En consecuencia, sistemas GIR deben ser capaces de resolver información implícita contenida en los documentos y en las consultas para poder entregar una respuesta apropiada a las consultas geográficas.

En el estado actual del área de recuperación de información geográfica [34, 88, 45, 69] ha sido posible observar que máquinas tradicionales de IR son capaces de recuperar un porcentaje considerable de documentos relevantes para la mayor parte de las consultas geográficas (alrededor del 80 %), sin embargo, presentan dificultades al momento de generar un orden pertinente de los documentos recuperados, lo cual resulta en un desempeño deficiente. Una de las razones por las que el ordenamiento es inapropiado es la falta de información en la consulta. Es por esta razón que grupos de investigación han tratado de cubrir esta falta de información empleando robustos recursos geográficos (e.g. ontologías geográficas), mientras que algunos otros grupos de investigación lo han intentado por medio de estrategias de reformulación de consultas a través de técnicas de retroalimentación de relevancia (Véase el capítulo 3).

En este trabajo proponemos una estrategia re-ordenamiento[11] para sistemas GIR. Recordemos que el problema de IR tiene dos grandes sub-problemas, primero determinar qué documentos son *relevantes*, y segundo determinar el *orden por relevancia* de los documentos recuperados. Aunque el recuperar documentos relevantes a consultas geográficas no es un problema que este completamente resuelto, nuestro trabajo parte de que la máquina de IR ha logrado recuperar un subconjunto de documentos relevantes, y nos enfocamos únicamente en el proceso de mejorar el *ordenamiento por relevancia* asignado a los documentos, con la finalidad de mejorar la calidad de la salida del sistema de recuperación. Adicionalmente, dado que las consultas geográficas tienden a incluir información implícita, nosotros proponemos el uso de documentos completos en lugar de palabras aisladas en el proceso del re-ordenamiento, a los cuales denominamos *documentos ejemplo*, con el propósito de hacer una aproximación de esta información implícita a una forma más explícita.

[11] En Inglés se le conoce como *re-ranking* o *ranking refinement*

## 1.4. Objetivos

Dados estos antecedentes, definimos el objetivo general de la tesis como:

*"Proponer un método de re-ordenamiento por relevancia que utilice la información tanto temática como geográfica contenida en un conjunto de documentos ejemplo para re-ordenar la salida otorgada por una máquina de recuperación de información en el contexto de búsquedas geográficas."*

### 1.4.1. Objetivos Particulares

- Proponer una estrategia de re-ordenamiento considerando documentos ejemplo
- Determinar la forma de representación geográfica de los documentos que aporta mayor información a la estrategia de re-ordenamiento
- Evaluar la utilidad y complementariedad de la información temática y geográfica en la estrategia de re-ordenamiento

## 1.5. Organización de la Tesis

El documento está organizado de la siguiente forma: en el siguiente capítulo se presentan conceptos básicos para entender el contenido de esta tesis, los cuales incluyen principalmente conocimientos de recuperación de información. Además de esto se incluyen conceptos de generación automática de resúmenes y de campos aleatorios de Markov, donde se abordan conceptos clave que sirvieron a la arquitectura propuesta en el proceso del re-ordenamiento.

En el capítulo 3 se presenta una revisión del trabajo más relevante y reciente en el área de recuperación de información geográfica. En el capítulo 4 se resume la contribución de este trabajo, de manera particular se exhibe la idea así como la arquitectura propuesta para resolver el problema del re-ordenamiento. En el capítulo 5 se plantean los experimentos realizados así como los resultados obtenidos a cada uno de estos. Finalmente, en el capítulo 6 se resumen las principales aportaciones de la investigación así como las conclusiones, y discutimos posibles direcciones para el trabajo futuro.

# Capítulo 2

# Preliminares

El objetivo de este capítulo es describir e introducir al lector de manera rápida a la teoría que fundamenta el trabajo realizado en esta tesis. Se presentan varias definiciones formales utilizadas a lo largo del documento. En primer lugar se pretende familiarizar al lector con la tarea de Recuperación de Información (sección 2.1). Posteriormente, la sección 2.2 describe la tarea de generación automática de resúmenes y finalmente se introduce la teoría de los campos aleatorios de Markov (sección 2.3), siendo ambos conceptos importantes en el desarrollo de este trabajo de investigación.

## 2.1. Recuperación de Información

La tarea principal de los sistemas de recuperación de información (IR) consiste en dada una consulta, formulada en lenguaje natural por algún usuario, obtener documentos relevantes que satisfagan las necesidades de información del usuario [36].

En este contexto, el término *relevantes* se refiere a que los documentos recuperados deberán estar semánticamente relacionados a la necesidad de información del usuario. Además, los documentos deberán estar ordenados de acuerdo a la *relevancia* que cada uno de estos tenga con respecto a la consulta dada. Nótese que la relevancia sólo puede ser evaluada por el usuario que formula la consulta, por lo que la evaluación de la efectividad de un sistema de recuperación de información no puede ser del todo objetiva[1]; aunque actualmente se han creado colecciones de documentos estándar y desarrollado protocolos para la evaluación y comparación de este tipo de sistemas [1, 2, 7, 36]. Nótese también que no se especifica el tipo de colección ni la modalidad de los

[1] A menos que conozcamos los documentos relevantes para cada posible petición de cada usuario de la colección.

documentos, por lo que por documento nos referiremos indistintamente a documentos de texto, imágenes, audio, etc.

### 2.1.1. Modelos de Recuperación de Información

Para solucionar la tarea de recuperación de información se han propuesto muchos modelos de recuperación. Un modelo de recuperación es el conjunto de métodos y estrategias que nos permiten representar y organizar la colección de documentos, definir consultas y compararlas con los documentos en la colección. En la siguiente sección se describe uno de los principales modelos de recuperación que se han propuesto. El lector con conocimientos básicos en recuperación de información puede omitir la presente sección, puesto que únicamente presenta los fundamentos de la tarea de recuperación y una breve descripción del principal modelo propuesto para esta tarea.

#### Modelo de espacio vectorial

El modelo de espacio vectorial (*VSM*, por sus siglas en Inglés), propuesto por Salton et al [29] se basa en la idea de que el significado de un documento está dado por las palabras que éste contiene. Propone llevar los documentos y la consulta a una representación vectorial, obtenida por las palabras contenidas en los documentos y consulta, donde la comparación de los vectores nos indique la similitud entre consulta y documentos.

En el *VSM* cada documento $d$ es representado por un vector ($\vec{d_i}$) de longitud igual al tamaño del vocabulario $|V|$. El vocabulario de la colección es el conjunto de todos los términos (e.g., palabras) diferentes que ocurren en la colección. Cada elemento $j$ del vector $\vec{d_i}$ indica la contribución del término $j$ en el documento representado por el vector $\vec{d_i}$. El conjunto de vectores que representan a los documentos contenidos en la colección generan un espacio vectorial donde los documentos pueden ser comparados a través de sus representaciones. Este espacio vectorial se representa por la matriz ($M^{TD}$) , denominada término-documento (TD), de dimensiones $N \times M$, donde $N$ es el tamaño del vocabulario de la colección, $N = |V|$, y $M$ es el número de documentos en la colección. Cada entrada $M^{TD}_{i,j}$ indica el peso o contribución del término $t_j$ en el documento $d_i$.

Diversos esquemas de pesado han sido propuestos, aunque el más utilizado es el denominado *tf-idf* (por, *term-frequency inverse-document-frequency*). La forma en

que se determina el valor de cada entrada $M_{i,j}^{TD}$ se muestra en la fórmula (2.1.1).

$$M_{ij}^{TD} = tf_{ij} \times log(\frac{|D|}{df_j}) \tag{2.1.1}$$

donde $tf_{ij}$ indica el número de ocurrencias del término $j$ en el documento $d_i$, $|D|$ es el número total de documentos en la colección y $df_j$ es el número de documentos que contienen el término $j$.

Las consultas en el *VSM* son especificadas por sentencias de texto que son consideradas un documento. Este documento es transformado a la misma representación vectorial que la colección. Una vez que consulta y documentos se encuentran en la misma representación es posible comparar el vector que representa la consulta y cada uno de los vectores en la colección. La medida de similitud más utilizada en el *VSM* es la denominada medida del coseno, descrita en la fórmula 2.1.2,

$$sim(q, d_i) = \frac{\sum_{j=1}^{|q|} w_{qj} d_{ij}}{\sqrt{\sum_{j=1}^{|q|}(d_{ij})^2 \sum_{j=1}^{|q|}(w_{qj})^2}} \tag{2.1.2}$$

donde $sim(q, d_i)$ indica la similitud entre la consulta $(q)$ y el documento $d_i$. $w_q$ es el vector construido con los términos contenidos en la consulta $q$ y $d_i$ es el vector que representa al documento $d$. Con esta fórmula estamos midiendo el ángulo en el espacio $|V|$ dimensional entre dos vectores; considerando una normalización para que la magnitud de los vectores no afecte el proceso de recuperación. Los documentos más cercanos a la consulta son regresados como relevantes. Algunas otras medidas de similitud que han sido propuestas son el coeficiente DICE y la medida JACCARD, descritos en las fórmulas 2.1.3 y 2.1.4 respectivamente.

$$sim(q, d_i) = \frac{2\sum_{j=1}^{|q|} w_{qj} d_{ij}}{\sum_{j=1}^{|q|}(d_{ij})^2 + \sum_{j=1}^{|q|}(w_{qj})^2} \tag{2.1.3}$$

$$sim(q, d_i) = \frac{\sum_{j=1}^{|q|} w_{qj} d_{ij}}{\sum_{j=1}^{|q|}(d_{ij})^2 + \sum_{j=1}^{|q|}(w_{qj})^2 - \sum_{j=1}^{|q|} d_{ij} w_{qj}} \tag{2.1.4}$$

Para los fines de este trabajo sólo se consideró la medida del coseno como forma de medir la similitud para la realización de los experimentos, para ver más detalles sobre éstas y algunas otras medidas de similitud, refiérase a [7, 36].

### 2.1.2. Evaluación de un Sistema de Recuperación de Información

En la Figura 2.1 se muestran las diferentes categorías en las que se clasifica a los documentos para cualquier consulta. En la figura es posible observar que existen documentos *recuperados* y documentos que son *relevantes*. En un sistema perfecto, esos dos conjuntos deberán ser iguales, i.e., sólo se recuperaran documentos relevantes. Aunque en la realidad, los sistemas recuperan muchos documentos que no son relevantes. Para medir la efectividad, existen dos medidas comunes: *precisión* y *recuerdo*. La precisión es la razón del número de documentos relevantes recuperados entre el total de documentos recuperados. La precisión es un indicador de la calidad del conjunto de documentos entregados como respuesta a la consulta del usuario. Sin embargo, esto no considera el número total de documentos relevantes. Un sistema podrá tener una muy buena precisión recuperando diez documentos de los cuales nueve sean relevantes (i.e., 0.9 de precisión), pero también es necesario tomar en cuenta el número de documentos relevantes. Si solamente hubiera nueve documentos relevantes, el sistema tendrá en efecto mucho éxito en el desempeño de su tarea, sin embargo si existieran millones de documentos relevantes, éste no seria tan buen resultado.


**Figura 2.1: Conjunto de Documentos Recuperados, Relevantes y Relevantes Recuperados**

El recuerdo considera el número total de documentos relevantes; es la razón del número de documentos relevantes recuperados entre el número total de documentos relevantes existentes en la colección. El cálculo del número de documentos relevantes, no es una tarea trivial. La única forma de hacerlo es revisar en su totalidad la colección

de documentos.

La *precisión* puede ser calculada a varios niveles de *recuerdo*. Considere una consulta $q$. Sabemos que para $q$ existen dos documentos relevantes. Cuando el usuario introduce $q$ al sistema de recuperación de información, éste le devuelve diez documentos, incluyendo los dos relevantes. Los documentos relevantes aparecen en la posición dos y cinco de la lista de documentos recuperados. La línea inclinada de la Figura 2.2 muestra que después de recuperar dos documentos se han encontrado sólo un documento relevante, es decir se ha alcanzado un cincuenta por ciento de *recuerdo*. En este mismo punto se tiene igual porcentaje de *precisión*, pues de los dos documentos recuperados sólo uno es relevante.

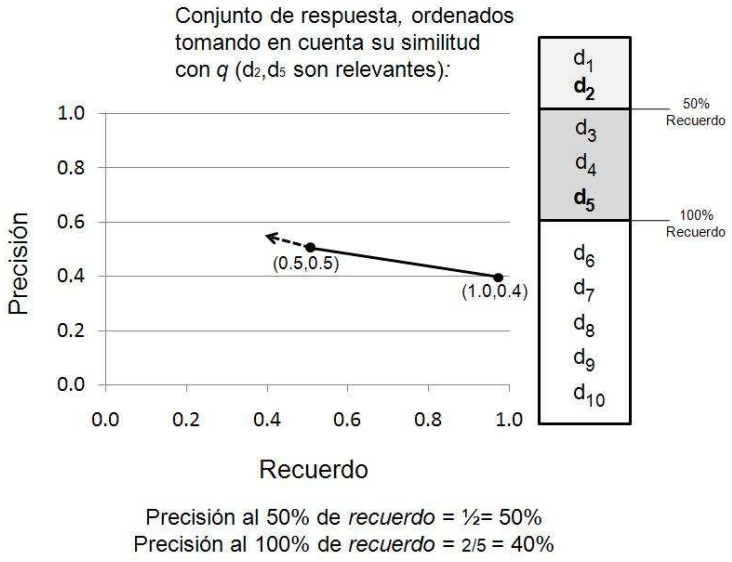


**Figura 2.2: Precisión y dos puntos de Recuerdo**

Así, para alcanzar 100 % de recuerdo, es necesario seguir evaluando el resto de los documentos recuperados. Para el ejemplo basta con revisar hasta los primeros cinco documentos recuperados. En este punto la precisión es del 40 %, ya que sólo dos de los cinco documentos recuperados son relevantes.

El comportamiento típico de la curva *Recuerdo/Precisión* puede ser observado en la Figura 2.3. Mientras mayor *recuerdo* se desea obtener, mayor cantidad de documentos deben ser recuperados. En un sistema ideal solamente los documentos relevantes serán recuperados. Lo cual significara que a cualquier nivel de *recuerdo* la *precisión* será siempre de 1.0.

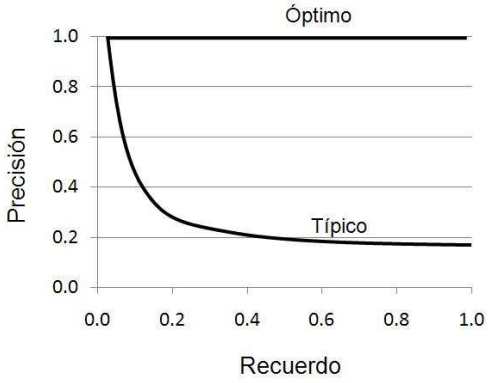


**Figura 2.3: Curva Recuerdo y Precisión (Típica y óptima)**

La *Precisión Promedio* o *Average Precision* (AveP) de la consulta $q$, descrita en la fórmula 2.1.5, se refiere al promedio de las precisiones calculadas en varios niveles de recuerdo para una consulta dada.

$$AveP(q) = \frac{\sum_{r=1}^{Ret_{docs}} P(r) \times rel(r)}{Rel_{docs}} \tag{2.1.5}$$

donde $P(r)$ es la precisión del sistema a los $r$-documentos considerados y $rel(r)$ es una función binaria que nos indica si el documento recuperado $r$ es relevante a la consulta o no. Cuando sucede que ningún documento relevante es recuperado para la consulta $q$, el valor de la precisión promedio es 0. El $AveP$ es una medida aproximada del área bajo la curva Recuerdo/Precisión (Figura 2.3). Intuitivamente, esta medida indica que tan bien el sistema regresa documentos relevantes en las primeras posiciones.

En muchos foros actuales encargados de evaluar el desempeño de los sistemas de recuperación de información, se hace uso de colecciones cerradas de documentos. Lo cual permite a los diferentes grupos de investigación reportar sus resultados en términos del $AveP$ y de la medida $MAP$. Donde el MAP sólo es la media aritmética de las $AveP$ obtenidas para cada consulta [7].

Dado que para muchos sistemas es importante saber cuantos documentos relevantes son obtenidos en las primeras posiciones, otra medida común entre los sistemas de IR es la *Precisión a k* ($P@k$), la cual indica la precisión del sistema IR a los $k$ documentos recuperados. Adicionalmente, también se emplea la medida *R-prec*, la cual se define como la precisión del sistema alcanzada después de haber recuperado $R$ documentos, donde $R$ es el número total de documentos relevantes que existen dentro

de la colección para la consulta $q$.

### 2.1.3. Retroalimentación de Relevancia

El proceso de retroalimentación de relevancia o *relevance feedback* (en Inglés) es una técnica controlada para la reformulación de consultas, donde la idea básica consiste en escoger ciertas palabras relacionadas a un conjunto de documentos previamente recuperados los cuales han sido identificados como relevantes por un usuario, de tal forma que estas palabras son agregadas a la consulta original [28]. Cuando se requiere aplicar un proceso de retroalimentación de relevancia, es necesario que se lleve a cabo una primera búsqueda, la cual producirá una lista ordenada de documentos, posteriormente el usuario inspecciona los primeros documentos recuperados para establecer su relevancia con respecto a la consulta dada. Estos juicios de relevancia otorgados por el usuario permiten al sistema calcular un nuevo conjunto de valores que reflejaran de manera más precisa la importancia de cada una de las palabras de la consulta.

Existen dos mecanismos bien definidos para aplicar un proceso de retroalimentación de relevancia los cuales son:

- Retroalimentación Manual o Simulada: documentos relevantes son identificados por un usuario y el nuevo conjunto de palabras que conformarán la consulta reformulada son seleccionados ya sea de forma manual o a través de un proceso automático.

- Retroalimentación Ciega: los documentos relevantes son identificados de manera automática por medio de suponer que los documentos colocados al principio de la lista son en efecto documentos relevantes y generalmente las nuevas palabras que conformarán la consulta reformulada son seleccionadas de manera automática.

Típicamente la selección de las palabras que conformarán a la consulta reformulada es por medio de considerar a las $n$ palabras más frecuentes dentro de los juicios de relevancia otorgados ya sea por un esquema de retroalimentación simulado o ciego. Entre las ventajas de aplicar un proceso de retroalimentación de relevancia es que generalmente es posible incrementar la precisión del sistema de IR siempre y cuando se incremente el tamaño de la consulta con buenas palabras, i.e., palabras relevantes a la necesidad de información del usuario. Entre las desventajas de este proceso está el

hecho de que implica un mayor costo computacional, dado que es necesario efectuar un segundo proceso de recuperación. Además, es una técnica muy sensible a la calidad de las palabras agregadas, dado que agregar una palabra muy mala puede afectar fácilmente el desempeño del sistema.

## 2.2. Generación Automática de Resúmenes

Un resumen generado automáticamente, comúnmente denominado "sumario", lo podemos definir de la siguiente manera:

- **Definición:** Un resumen o sumario es un texto producido a partir de uno o más documentos, que contiene toda o la mayor parte de la información más significante contenida en los documentos originales, y además su extensión es significativamente menor a la de los documentos originales [42, 43].

Dentro del área de generación automática de resúmenes existen dos formas básicas en las cuales los resúmenes son generados:

- Los **resúmenes basados en extractos** son los resúmenes que son creados al reutilizar porciones (palabras, oraciones, párrafos, etc.) de los documentos origen, entregando al usuario como resumen final estas porciones ordenadas de alguna forma en específico[2].

- El otro tipo de resúmenes denominados simplemente como **resúmenes** ó **abstractos** son los que *generan* el resumen final a partir de los extractos obtenidos en un primer paso. El proceso de *generar* un resumen involucra crear nuevas oraciones a partir de las que han sido identificadas como relevantes. Para poder crear estas nuevas oraciones es necesario contar con sofisticados recursos lingüísticos que interpreten adecuadamente contenido y significado de las oraciones extraídas. Una vez hecha esta interpretación, el sistema puede mezclar y/o comprimir oraciones con el objetivo de entregar al usuario un resumen más coherente.

[2] A lo largo del documento de tesis se hará uso de la palabra *"resumen o resúmenes"* para referirnos a un *"resumen basado en extractos"* a no ser que explícitamente se exprese lo contrario.

### 2.2.1. Generación de Resúmenes de Múltiples Documentos

La tarea de realizar resúmenes de un solo documento ya es muy compleja por si sola. Pero la tarea de realizar el resumen de una colección de documentos relacionados temáticamente posee varios retos adicionales. El proceso de generar un resumen de múltiples documentos consiste en la creación de un resumen simple a partir de conjunto de documentos relacionados temáticamente. Existen tres grandes problemas que surgen al momento de manejar múltiples documentos: (i) reconocer y resolver redundancias, (ii) identificar diferencias importantes entre los documentos, y (iii) asegurar la coherencia del resumen, tomando en cuenta que diferentes porciones de información provienen de diferentes fuentes.

Entre los métodos propuestos para resolver el problema de generar resúmenes de múltiples documentos se propone el uso de técnicas de extracción de información[3] para facilitar el proceso de identificación de similitudes y diferencias entre documentos. Sin embargo, una de las principales desventajas de este tipo de técnicas es que los resúmenes generados tienden a contener únicamente ciertos tipos de información predefinida, i.e., contendrían únicamente la información que le interesa y para la cual está entrenado el sistema de extracción de información.

Comúnmente se utilizan medidas de similitud para lograr la identificación de redundancias en los documentos. Una técnica muy común consiste en medir la similitud entre cada par de oraciones y posteriormente un proceso de agrupamiento es aplicado con la finalidad de encontrar los temas comunes dentro de la colección de documentos [86]. Una vez hecha la identificación de la información similar, esta debe de ser incluida en el resumen. Más allá de simplemente listar estas porciones de información (como se hace al trabajar con un documento), es necesario seleccionar las oraciones más representativas asegurando así que la mayor cantidad de información esta siendo considerada para el resumen.

Asegurar la coherencia en el resumen final es una tarea difícil, pues en principio se requiere de un cierto nivel de entendimiento del contenido de cada oración además de conocimiento sobre reglas del discurso. Debido a la complejidad que esto involucra, actualmente muchos de los sistemas simplemente siguen la línea del tiempo para

[3]Los sistemas de extracción de información (IE) realizan la tarea de buscar información muy concreta en colecciones de documentos, detectar la información relevante, extraerla y presentarla en un formato estructurado. Generalmente la información extraída es almacenada en bases de datos, e.g., bases de datos sobre desastres naturales.

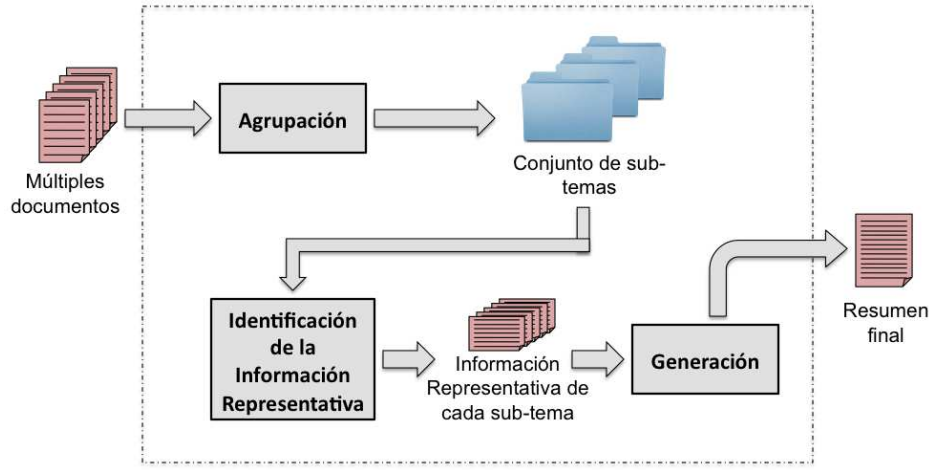


**Figura 2.4: Arquitectura general de un sistema generador de resúmenes de múltiples documentos**

ordenar el contenido del resumen y en algunos casos una etiqueta indicando la fecha del evento acompaña a cada oración.

La figura 2.4 muestra de manera gráfica la arquitectura general de un sistema generador de resúmenes de múltiples documentos.

Dentro de esta arquitectura general se puede observar los siguientes pasos generales a seguir:

- **Agrupación.** El proceso de agrupación tiene como principal objetivo permitir al sistema dividir la colección inicial en sus diferentes sub-temas. Esto permite identificar las similitudes entre documentos además de la información única dentro de cada uno de ellos.

- **Identificación de información representativa.** Una vez que los diferentes sub-temas han sido clasificados, es necesario contar con un método de selección de los elementos más representativos de cada grupo. El objetivo es abarcar la mayor cantidad de información posible con el menor número de oraciones.

- **Generación.** Al igual que en la tarea de crear el resumen de un solo documento, el proceso de generación consiste en crear un documento coherente. En la mayoría de los casos la etapa de generación no es alcanzada, las oraciones son ordenadas de acuerdo a la línea del tiempo.

Finalmente, una vez generado el resumen viene el problema de cómo saber si es un resumen "bueno". La tarea de evaluar la calidad de los resúmenes producidos de manera automática siempre ha sido un problema para los grupos de investigación dentro de esta área. La evaluación de un resumen se vuelve una tarea muy subjetiva debido a que no existe un resumen ideal contra el cual se pueda comparar la salida de un sistema generador de resúmenes. Sin embargo, existen dos tipos de evaluación que comúnmente han sido empleados, evaluación **intrínseca** que mide la calidad de la salida y **extrínseca** que miden la ayuda o asistencia que la salida proporciona al usuario en el desempeño de una tarea particular.

Como se verá en el capitulo 4, nuestro objetivo al generar resúmenes es conservar sólo la información importante de un grupo de documentos relacionados temáticamente y posteriormente emplear el resumen generado en el proceso de re-ordenamiento propuesto en esta tesis. En este sentido, estaremos aplicando un tipo de evaluación extrínseca para determinar la calidad de los resúmenes generados.

## 2.3. Campos de Markov

Los Campos Aleatorios de Markov (MRF, por sus siglas en Inglés) pertenecientes a la teoría de probabilidad son una herramienta útil para analizar dependencias espaciales o contextuales de fenómenos físicos. En otras palabras, un MRF es un modelo gráfico probabilista que combina conocimiento *a priori* dado por observaciones con conocimiento obtenido de las interacciones de las variables con sus vecinas [60].

El concepto de MRF viene del intento de colocar dentro de un marco probabilista general un modelo físico específico conocido como modelo de Ising. En este modelo se trataba de explicar hechos empíricos observados en materiales ferromagnéticos. Los MRF son una extensión de las cadenas de Markov, en los cuales el índice del tiempo se sustituye por el índice espacial [18].

La primera formulación del modelo de Ising es la siguiente: considere una secuencia $0, 1, 2, 3, ..., n$ de puntos en una línea. En cada punto o sitio existe un pequeño dipolo que en cualquier momento puede tomar uno de los valores + o −. El modelo le asigna una medida de probabilidad a cada una de las posibles configuraciones de valores. El valor de cada dipolo está influenciado por el valor de sus dipolos vecinos. A esta medida se le conoce como *Campo Aleatorio* [49]. La figura 2.5 muestra gráficamente la formulación del modelo de Ising. Existen dos factores que determinan la probabilidad

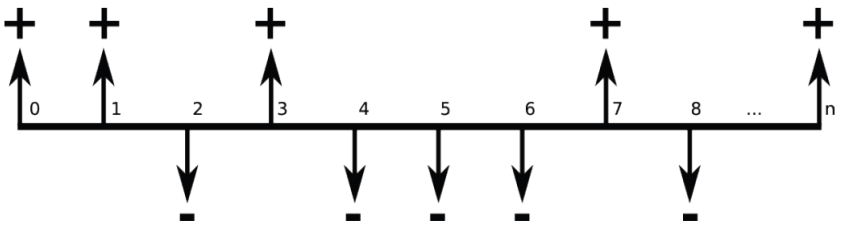


**Figura 2.5: Esquema del modelo de Ising**

de una configuración de valores. La primera es la probabilidad *a priori* de cada estado, que se ejemplifica con un campo magnético externo. La segunda es la probabilidad conjunta, representada por la intersección de los campos magnéticos de los dipolos vecinos.

Formalmente se puede definir un MRF de la siguiente forma; sean $F = \{F_1, F_2, ..., F_n\}$ variables aleatorias dentro de un conjunto $S$, donde cada $F_i$ puede tomar un valor $f_i$ de un conjunto de valores $L$. A $F$ se le conoce como un campo aleatorio, y a la instanciación de cada una de las variables $F_i \in F$ con un valor $f_i$, se le llama configuración de $F$, por lo tanto, la probabilidad de que una variable aleatoria $F_i$ tome el valor $f_i$ se denota como $P(f_i)$, y la probabilidad conjunta es denotada como $P(F_1 = f_1, F_2 = f_2, ..., F_n = f_n)$.

Se dice que un campo aleatorio es un campo aleatorio de Markov si éste tiene la propiedad de *localidad*, es decir que el campo satisfaga la siguiente propiedad:

$$P(f_i \mid f_{S-\{i\}}) = P(f_i \mid f_{N_i}) \tag{2.3.1}$$

donde $S - \{i\}$ representa el conjunto de $S$ sin el $i^{esimo}$ elemento, $f_{N_i} = \{f_{i'} \mid i' \in N_i\}$, y $N_i$ representan el conjunto de variables vecinas del nodo $f_i$.

Un sistema de vecindad para $S$ se define como:

$$V = \{V_i \mid \forall i \in S\} \tag{2.3.2}$$

el cual cumple las siguientes propiedades: *i)* un sitio no es vecino de si mismo, y *ii)* la relación de vecindad es mutua.

La probabilidad conjunta puede expresarse como:

$$P(f) = \frac{e^{-U_p(f)}}{Z} \tag{2.3.3}$$

donde $Z$ es conocida como la función de partición o constante de normalización, y $U_p(f)$ es conocida como la función de energía. La función de energía $U_p(f)$ representa la información externa e interna necesaria para decidir cambiar o no el valor de una variable aleatoria.

La configuración óptima es obtenida cuando se minimiza la función de energía $U_p(f)$ obteniendo un valor para cada una de las variables aleatorias en $F$. Obtener la configuración de menor energía (mayor probabilidad) es una operación muy costosa, por lo que se plantea como un problema de optimización. Es decir que se busca la configuración de mayor probabilidad, sin tener que calcular directamente las probabilidades de cada configuración.

Para plantear la obtención de la configuración más probable como un problema de optimización, se necesitan definir 3 componentes principales [18]:

**1**.- La representación del MRF.- se representa al MRF incluyendo su sistema de vecindad y los potenciales asociados a los factores de probabilidad.

**2**.- La función objetivo.- se define a la función de energía incluyendo los potenciales definidos en el paso anterior, los cuales son los que se quiere minimizar.

**3**.- El algoritmo de optimización.- se selecciona un algoritmo que permita seleccionar el valor más apropiado para las variables en $F$. Los pasos básicos para encontrar la configuración más probable son los que se muestran a continuación:

```
1.- Inicializar cada variable del MRF con un valor aleatorio
2.- Para cada variable del MRF hacer:
      ▪ Calcular el valor de energía de cada variable con base
        en la función de energía
      ▪ De acuerdo a un algoritmo de optimización, seleccionar
        el posible valor para la variable
3.- Repetir el paso 2 hasta que el MRF converja a la
    configuración más probable o se cumplan N iteraciones
4.- Obtener la configuración óptima
```

Los algoritmos de optimización tienen como función el seleccionar el valor adecuado durante la iteración del algoritmo. De acuerdo a esa característica, los algoritmos

usados comúnmente son tres. *i)* el algoritmo Metrópolis, el cual selecciona con una probabilidad fija el estado con mayor energía. *ii)* el Recocido Simulado, el cual a lo largo de la ejecución del algoritmo va decrementando la probabilidad de seleccionar el estado de mayor energía. Finalmente, *iii)* el algoritmo ICM (Iterated Conditional Modes por sus siglas en Inglés), el cual toma siempre el estado de menor energía [18, 60].

Así entonces, la idea principal de todo esto (Véase capítulo 4) es integrar mediante el uso de un MRF, el orden original obtenido por la máquina de IR, la similitud entre los documentos y los *documentos ejemplos*, con la finalidad de generar un orden más apropiado para la lista de documentos recuperados por el sistema GIR. El MRF deberá separar los documentos relevantes de los que no lo son mediante el uso de una función de energía, la cual deberá representar la similitud entre los documentos de la misma clase haciendo uso de algún conjunto de características. Además, la función de energía también debe representar la información *a priori* de los documentos, la cual puede ser representada como la similitud con la consulta o tomando en cuenta el orden original generado por la máquina de IR. Finalmente, el uso de los *documentos ejemplo* permitirá incluir la intención de búsqueda en la función de energía.

# Capítulo 3

# Trabajo Relacionado

El presente capítulo tiene como objetivo ubicar al lector en el área donde se enfoca el trabajo de tesis, i.e., en el desarrollo de sistemas para la recuperación de información geográfica. Durante la revisión del trabajo relacionado se darán a conocer los principales problemas a los que sistemas GIR se enfrentan actualmente. Se describirá una arquitectura general de un sistema GIR, propuesta en este trabajo como tal; donde tomando en cuenta los módulos que en ella se mencionan se describirán las tendencias que diferentes grupos de investigación han adoptado para la resolución de los problemas existentes en cada uno de estos módulos.

## 3.1. Recuperación de Información Geográfica

Sabemos que idealmente cualquier sistema de recuperación de información debería satisfacer las necesidades de información de los usuarios, y no solamente procesar una consulta dada. Un sistema de Recuperación de Información Geográfica (GIR), con la finalidad de responder a consultas geográficas[1], debería entonces ser capaz de manejar las consultas de acuerdo a las necesidades de información implícitas en ellas [40].

Mucha de la investigación que se ha hecho trata de atacar el problema desde varios puntos de vista. Existen los trabajos que se enfrentan al problema de identificar de manera automática una consulta geográfica [35, 40, 68, 78], con la finalidad de definir la mejor forma de manejarla dentro del sistema de recuperación de información. Algunos otros trabajos han enfocado sus esfuerzos en examinar las expresiones en lenguaje natural contenidas en los documentos con la finalidad de asignar una etiqueta geográfica a dichos documentos [6, 8, 20, 22, 41, 81, 73, 80, 89], i.e., crear colecciones

[1]Vea Sección 1.2 para entender como se define una consulta geográfica

de documentos *"geo-etiquetados"*; a partir de esto, definen estrategias de búsqueda especiales. Así pues, cuando en una consulta se identifica algún término geográfico, éste es mapeado a su correspondiente(s) etiqueta(s) geográfica(s). Por ejemplo, el término "Madrid" puede tener como etiquetas geográficas a: España, Europa, Peninsula Ibérica, etc. De esa forma, busca asegurarse que los documentos recuperados contengan elementos geográficos relacionados directamente con los términos de la consulta, en otras palabras, que tengan el mismo *foco* geográfico. Uno de los problemas importantes con los que se enfrentan estos trabajos es la desambiguación de términos geográficos, que como se puede ver en [10, 59, 61, 72] es por si solo un problema complicado. Por ejemplo, para el término *Córdoba* pueden existir al menos tres posibles etiquetas geográficas que son: "España", "México" y "Argentina".

Por otra parte, encontramos algunos trabajos [14, 27, 59, 72] que buscan por relaciones semánticas entre las consultas y bases de datos geográficas (GIS), o entre consultas y ontologías geográficas, de forma que les es posible proponer un esquema de expansión de consultas. De esta forma, si se logra hacer una adecuada expansión geográfica, se espera que una máquina de IR tradicional pueda ser capaz de resolver la necesidad de información de los usuarios.

Finalmente, algunos grupos de investigación proponen nuevas métricas para hacer el ordenamiento de los documentos que la máquina de IR obtuvo [5, 13, 55, 66, 67, 91]. En una máquina de recuperación de información tradicional, los documentos entregados al usuario son ordenados internamente por la máquina de IR basándose en el grado de similitud que tienen con la consulta dada. Sin embargo, estos grupos de investigación afirman que en el contexto de búsquedas geográficas, este ordenamiento es inadecuado. Así pues, lo que proponen es tener un módulo de ordenamiento alterno, el cual considere de manera especial los términos geográficos contenidos tanto en la consulta como en los documentos recuperados. Un factor común dentro de estos trabajos es la necesidad de contar con *"geo-etiquetas"* para poder medir la cercanía de los términos geográficos y de esta forma ir ordenando los documentos recuperados de acuerdo a su similitud o cercanía geográfica.

En las siguientes secciones se expone con mayor detalle algunos de los enfoques empleados actualmente para tratar de resolver los problemas que hemos mencionado.

## 3.2. Componentes de un Sistema GIR

En la Figura 3.1 se muestra una arquitectura GIR básica. Es conveniente mencionar que esta arquitectura no es utilizada en forma general. Los diferentes módulos que se muestran no son siempre utilizados en conjunto, sin embargo, con la finalidad de hacer más fácil la lectura y ubicación de los diferentes problemas presentes en un sistema GIR es que se propone esta arquitectura como general.

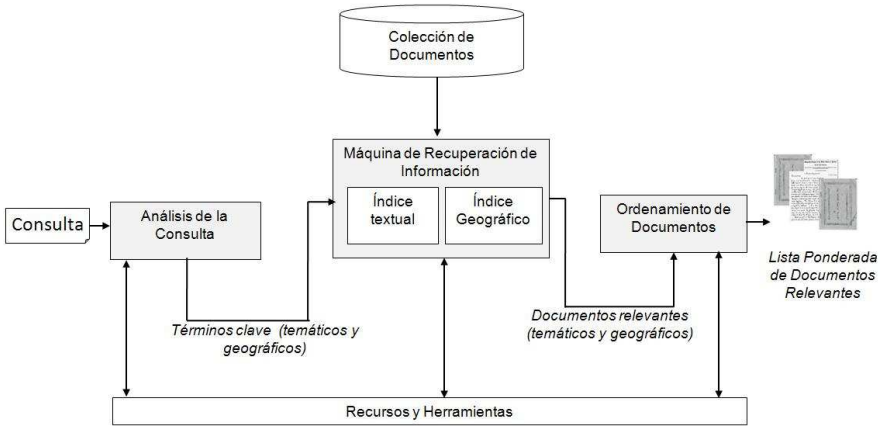


**Figura 3.1: Arquitectura general de un sistema GIR**

En las secciones siguientes se describirán tanto los problemas presentes en cada uno de los módulos mostrados en la figura 3.1, como las soluciones propuestas por diferentes grupos de investigación a dichos problemas.

### 3.2.1. Análisis de la Consulta

El primer módulo, *análisis de la consulta*, tiene por objetivo extraer de manera automática los términos clave de la consulta. Generalmente, estos términos son clasificados en *términos temáticos* y *términos geográficos*. Por ejemplo, en la consulta "Coches bomba en Madrid", la clasificación quedaría como: *temáticos: coches bomba*, *geográficos: Madrid*. Una de las formas más comunes para resolver el problema de identificar los términos geográficos contenidos en la consulta es por medio de un identificador de entidades nombradas (NER[2]), algunos trabajos que siguen esta estrategia son [13, 25, 30, 44, 50, 66, 82, 34].

[2]Name Entity Recognition

Nótese que el identificador de entidades nombradas sólo reconocerá los términos geográficos, sin embargo, el problema de la ambigüedad geográfica seguirá presente. Algunos grupos de investigación que han tratado de resolver este problema son [59, 61, 72]. Una de las técnicas utilizadas por estos grupos de investigación para lograr la desambiguación es por medio de métodos basados en reglas. El más común y sencillo de hacer es contar con un "gazetter básico" dentro del cual habrá una única entrada o localidad para cada una de las posibles entidades identificadas. Generalmente el "gazetter" contiene nombres de lugares más conocidos o "famosos".

Es importante tener en cuenta que el problema de ambigüedad no sólo se presenta entre lugares que tienen el mismo nombre, el problema de ambigüedad geográfica se da también con los nombres de organizaciones y/o nombres de personas. En ocasiones también sucede que los nombres de lugares se utilizan para referirse a entidades administrativas, por ejemplo, *habló con Washington*. Cuando una persona lee un documento que contiene nombres de lugares, éste es capaz de resolver cualquier posible ambigüedad ayudándose del conocimiento obtenido a través de las claves contextuales que aparecen dentro del mismo documento. La asignación automática de los "focos geográficos" y en consecuencia la desambiguación de términos geográficos intentan repetir los pasos empleados por las personas, es decir, consideran todos los términos geográficos que ocurren dentro del documento. Por ejemplo, si el nombre de un lugar ocurre en asociación con un conjunto de otros nombres, con los cuales comparte cierta vecindad, o son instancias con las cuales comparte la misma región padre, entonces esto da evidencia suficiente para distinguir el significado del término ambiguo [12].

Otra parte importante de la consulta es la *relación espacial*. Una relación espacial es una restricción que aparece en consultas geográficas la cual "selecciona" o "direcciona" una región geográfica. Este tipo de *selección* puede ser tan simple como "Lagos *en* el estado de Maine", o más complejas como "encuentra el camino *más corto* de Bostón a Bangor". Las relaciónes espaciales están divididas en tres categorias [34]: *i)* topológicas (e.g., museos en Roma); *ii)* métricas (e.g., no muy lejos de Roma); *iii)* proyectivas (e.g., a la derecha del camino).

Algunos trabajos que proponen la identificación y clasificación de la *relación espacial* son [30, 44, 68, 79, 82, 34]. Algunos de estos trabajos, en un enfoque sencillo, identifican la relación espacial y la eliminan del proceso de recuperación [30, 44, 82]. Por otro lado, en un enfoque más completo, se definen reglas[3] para cada uno de los

[3]Reglas definidas manualmente y que son distintas dependiendo el grupo de investigación

tipos de relaciones espaciales [68, 79, 34], por ejemplo, si la relación espacial contiene algo como: *"a 100km alrededor de Frankfurt"*, el primer paso es determinar que lugares son los que están alrededor de "Frankfurt", para lograr esto utilizan alguna base de datos geográficos. Posteriormente por medio de medir la distancia entre la ciudad de Frankfurt y cada uno de los lugares circundantes (todo esto se calcula en coordenadas espaciales), se determina cuáles de estos cumplen con la restricción espacial.

Para el caso de relaciones espaciales aún más ambiguas, como lo son *"at, in, near", etc.*, éstas siguen siendo convertidas a una cantidad numérica, la cual va a variar dependiendo de la importancia del lugar en cuestión. Por ejemplo, la definición de *"near"* será diferente si hablamos de la *cercanía* de dos países a que si se habla de la *cercanía* de dos poblaciones dentro de un mismo país.

En conclusión, dado que los términos de la relación espacial pueden ser tan ambiguos como lo son los términos geográficos contenidos en la consulta geográfica, los grupos de investigación que han enfocado sus esfuerzos en realizar un adecuado análisis de la consulta defienden la idea de que el interpretar adecuadamente estos términos permitirá realizar un adecuado proceso de recuperación de documentos en contextos geográficos.

### 3.2.2. Máquina de Recuperación de Información

Por otra parte, algunos grupos de investigación han propuesto que en el módulo de *recuperación de información* se haga uso de índices separados [67, 52, 44, 66, 13, 25, 17], i.e., un índice temático, el cual contiene todos aquellos términos que no correspondan a una referencia geográfica (*términos temáticos*), y uno o más índices geográficos, los cuales contienen sólo nombres de referencias geográficas (*términos geográficos*)[4].

Generalmente, el índice geográfico se construye con ayuda de diferentes recursos y/o herramientas. Ejemplos de recursos son la Web, catálogos u ontologías geográficas, etc. En [21] se muestra un conjunto de técnicas para la construcción de este tipo de recursos. Por otro lado, ejemplos de herramientas son identificadores de entidades nombradas (NER, por sus siglas en Inglés), técnicas de desambiguación, etc.

En una modalidad sencilla, [52] propone identificar por medio de una herramienta NER todas aquellas entidades geográficas contenidas en la colección de documen-

[4]En IR tradicional se maneja un único índice.

tos. Posteriormente, filtra con la ayuda de un catálogo geográfico aquellas entidades geográficas identificadas por su herramienta NER, conservando sólo aquellas que aparecen en el catálogo geográfico. Finalmente, con las entidades geográficas restantes construye el índice geográfico. Así entonces, cuando una consulta es dada, ésta se divide en parte *temática* y *geográfica*, mandando cada una a su índice correspondiente para aplicar un proceso de recuperación. En [13] un esquema similar es empleado, sólo que aquí no se construye un índice *temático*, en su lugar un se genera un índice *geográfico enriquecido*, el cual contiene holónimos y sinónimos de las entidades geográficas identificadas. Así, el proceso de recuperación se hace sobre ambos índices geográficos.

Es importante mencionar como nota final, que ninguno de estos dos trabajos [52, 13] aplica un proceso de desambiguación, en su lugar conservan todas las entidades que aparecen en el recurso, y si existen términos ambiguos conservan todas las posibilidades de dicho término existentes en el catálogo geográfico.

Por otro lado, esquemas más complejos [67, 44, 66, 17] construyen un índice geográfico que tiene la característica de poder expandir y asociar los diferentes términos geográficos encontrados en la colección de documentos. Por ejemplo, si en la colección de documentos se encontraron los términos *USA* y *Estados Unidos*, el índice geográfico verá a estos dos como uno mismo, de forma tal que si la consulta contiene el término *USA*, el índice geográfico permitirá a la máquina de IR recuperar aquellos documentos que contengan tanto el término *USA* como el término *Estados Unidos*.

En [67] además de poder asociar términos geográficos entre si, proponen también un método de expansión del índice geográfico, es decir, no importa si dentro de la colección sólo aparecieron los términos *USA* y *Estados Unidos*, pues por medio de una ontología geográfica es posible asociar estos términos con algunos otros, para nuestro ejemplo seria con términos como: *EE.UU., Norte América, América del Norte, etc*, e incluso les es posible asociar estos términos con los nombres de los estados del mismo país. Note que para poder hacer esto, todos estos trabajos aplican un proceso de desambiguación geográfica similar a los descritos en [12].

Normalmente, en una máquina de IR tradicional, una vez que se ha recuperado un conjunto de documentos, el mismo módulo de *IR* genera una lista ordenada de los documentos recuperados de acuerdo a su relevancia con la consulta dada por el usuario, la cual es generalmente entregada como salida del proceso de recuperación al usuario. Ahora bien, en un esquema con múltiples índices, se tendrá tras aplicar el

proceso de recuperación múltiples listas ordenadas de documentos recuperados. Por ejemplo, se tendrá una lista de documentos los cuales corresponden a todos aquellos documentos relevantes a los términos *temáticos* de la consulta, y por otra parte, se tendrá una o más listas las cuales corresponden a aquellos documentos relevantes a los términos *geográficos* de la consulta. En la siguiente sección (3.2.3) describiremos como este problema ha sido resuelto, de tal forma que al final se tenga una única lista ordenada de documentos recuperados.

En algunas ocasiones, como paso previo a entregar la lista final de documentos relevantes recuperados al usuario, se aplica la técnica conocida como *reformulación de consultas* o *expansión de consultas* a través de un proceso de retroalimentación de relevancia (Vea sección 2.1.3) . La cual consiste básicamente en incluir en la consulta términos relacionados (e.g., sinónimos) a los términos clave, con la finalidad de mejorar el desempeño del sistema de recuperación de documentos. Para el caso particular de un sistema GIR, la expansión de consultas se puede dar de las siguientes formas: *i)* aplicar la expansión en la parte *temática* de la consulta [23, 25, 56, 33, 38, 53, 66, 90, 88]; *ii)* aplicar la expansión en la parte *geográfica* de la consulta [4, 30, 50, 52, 67, 82, 90].

Como ejemplo considere la consulta "Coches bomba en Madrid", donde la parte temática correspondería a: "coches bomba"; una expansión temática intuitiva podría ser: "coches autos bomba explosión". Por otro lado, una posible expansión de la parte geográfica podría ser: "madrid barajas españa", mientras que una expansión en ambas partes de la consulta puede ser una combinación de estas dos. Nótese que para el caso de la expansión geográfica, ésta no se hace por medio de sinónimos, en su lugar, la expansión debe hacerse agregando términos relacionados geográficamente con el término original de la consulta [30, 67, 82].

Para el caso de la expansión en la parte temática de la consulta generalmente se sigue la estrategia de retroalimentación ciega (descrita en la sección 2.1.3), también conocida como "pseudo-retroalimentación de relevancia". Los diferentes grupos de investigación que han aplicado esta estrategia no coinciden en un número adecuado de documentos a considerar (desde 3 hasta 20 documentos), ni tampoco en el número adecuado de términos (desde 3 hasta 40 términos) a agregar en el proceso de la reformulación de la consulta. Por otro lado, un esquema más inteligente es mostrado en [90], donde la expansión de la consulta se hace por medio de agregar sinónimos y/o términos relacionados semánticamente a los que aparecen originalmente en la

consulta, los cuales son obtenidos de manera automática de una ontología. Un punto débil de este trabajo es que dicha ontología fue construida específicamente (*Ad hoc*) para el conjunto de consultas del GeoCLEF 2008[5].

Para el caso de la expansión en la parte geográfica podemos hablar de dos enfoques que hasta el momento se han seguido. Por un lado están los grupos que no realizan ningún tipo de desambiguación de los términos geográficos [52, 50, 82, 30, 90, 75], y por otra parte, los que si hacen desambiguación [4, 17].

Los trabajos que no realizan desambiguación (a excepción de [90]) en su mayoría realizan sólo un proceso de verificación, es decir, aquellas entidades geográficas identificadas por una herramienta NER son conservadas sólo si estas existen dentro de algún catalogo geográfico. Y la expansión generalmente se da agregando a la consulta toda o parte de la información que el catálogo geográfico contenga, por ejemplo, nombres de países circundantes, región o continente al que pertenece, capital, población, etc. Para el caso particular de [90], la expansión la realiza por medio de una ontología geográfica, aunque el punto clave de porqué en este trabajo no es necesaria una desambiguación es debido a que la ontología se construyó *Ad hoc* para el conjunto de consultas del GeoCLEF 2008.

Para el caso de los trabajos que si realizan un proceso de desambiguación [4, 17], el proceso de expansión geográfica es más "inteligente", pues el saber de manera específica de que región o lugar se habla en las consultas les permite agregar términos que efectivamente estén relacionados geográficamente con la consulta. En un esquema sencillo, en [4] agrega información de los términos desambiguados la cual es obtenida de un catálogo geográfico. Una de las características de este mismo trabajo es que define un conjunto de reglas para hacer la expansión dependiendo del tipo del término geográfico, es decir, si el término geográfico es una ciudad, expande hacia arriba, es decir país y continente al que la ciudad pertenece, mientras que si el término es un nombre de continente, expande hacia abajo, es decir agrega los nombres de los países contenidos en dicho continente. En [17] la expansión se hace de manera muy similar a lo propuesto por [4] sólo que aquí todo se hace por medio de una ontología geográfica la cual fue construida por varios recursos geográficos.

[5]El GeoCLEF es un foro de evaluación de sistemas GIR(http://www.clef-campaign.org/)

### 3.2.3. Ordenamiento por Relevancia

En el contexto de sistemas GIR, la necesidad de un módulo alterno de *ordenamiento* surge debido a la presencia de múltiples índices en el módulo de la máquina de recuperación de información, que como se mencionó en la sección anterior múltiples índices generan múltiples listas de documentos para las cuales el ordenarlas de manera adecuada se vuelve una tarea imprescindible.

Sin embargo, en trabajos más recientes [34, 45, 69, 88] el problema del ordenamiento es considerado debido a que se ha observado que máquinas de IR tradicionales son capaces de recuperar un porcentaje considerable de documentos relevantes para la mayor parte de las consultas geográficas (alrededor del 80 %), sin embargo, presentan dificultades al momento de generar un orden pertinente de los documentos recuperados, lo cual resulta en un desempeño deficiente. En otras palabras, máquinas de IR tradicionales obtienen un buen nivel de *recuerdo* pero una muy baja *precisión*[6]. Dentro de este trabajo de investigación, como se verá en el capítulo 4, y como se mencionó en la sección 1.3, nosotros también aprovechamos las ventajas que estas máquinas de IR tan robustas nos ofrecen, y nos enfocamos al problema de ordenar la salida otorgada por estos sistemas, es decir, en mejorar la *precisión* del sistema de recuperación.

Así entonces, el módulo de *"ordenamiento por relevancia"* (también conocido como: *relevance ranking* o *ranking refinement*) es aplicado a los documentos o conjuntos de documentos que son entregados por el módulo de recuperación de información (módulo de IR). Este módulo tiene como objetivo principal, tal y como su nombre lo dice, "ordenar" estos documentos. Para el caso particular de búsquedas geográficas, este ordenamiento debería realizarse no sólo tomando en cuenta la similitud temática de los documentos con las consultas, si no también tomando en cuenta una similitud o cercanía geográfica.

Durante los últimos años se han propuesto variados esquemas de ordenamiento para sistemas de información geográfica, desde aquellos que no involucran el uso de algún recurso geográfico, hasta aquellos que emplean varias medidas geográficas en el proceso del ordenamiento.

Como ejemplo de los primeros están los trabajos propuestos en [52, 24, 51, 83, 25, 9, 54, 87]. Una característica que tienen en común estos trabajos es que parten

[6]Vea sección 2.1.2 para conocer la definición de las medidas de *recuerdo* y *precisión*

de la existencia de dos o más listas de documentos recuperados por algún sistema de IR, el cual puede o no tener alguna configuración especial para procesar las consultas geográficas.

En el esquema más sencillo, en [24, 51, 83, 25] se propone el uso de operadores lógicos entre las distintas listas de documentos recuperados (e.g., AND). De esta forma, la lista final que se entregará al usuario como resultado está conformada por aquellos documentos que aparecieron en todas las listas a combinar. Por ejemplo, suponiendo una máquina de IR con dos índices, la lista final se formará a partir de aquellos documentos que fueron recuperados por el índice temático y también por aquellos que fueron recuperados por el índice geográfico. Como paso final, después de haber aplicado el operador lógico a las listas de documentos, la lista resultante es ordenada tomando en cuenta el *rank*[7] que la máquina de IR otorgó a cada documento.

Siguiendo con los trabajos que no emplean recursos geográficos para aplicar un ordenamiento a los conjuntos de documentos recuperados tenemos a [52, 9, 54, 87]. La característica en común de estos trabajos es que aplican estrategias más inteligentes para hacer la mezcla de las listas de documentos recuperados por alguna máquina de IR. En particular estos trabajos emplean técnicas basadas en redundancia de información para hacer la construcción y ordenamiento de la lista final. Como ejemplo de estas técnicas tenemos el conocido CombMNZ, que es una técnica tradicional dentro del área de fusión de información [57, 64].

En general estos trabajos tienen la ventaja de no depender de recursos externos para la realización de su tarea, sin embargo, para tener un buen desempeño dependen mucho de la calidad de las listas que van a fusionar, esto es debido a que si no hay la suficiente redundancia entre las listas estos métodos proporcionarán como salida una lista poco precisa, resultando en un desempeño deficiente de la máquina de IR.

Por otro lado, técnicas un poco más elaboradas que incluyen el uso de recursos geográficos son las de *filtrado de documentos* [39, 25, 74, 71, 34]. La idea básica detrás de estos sistemas es una vez que se tiene una lista de documentos recuperados, sobre los cuales de alguna forma se sabe cuál es su foco geográfico, se *filtran* dejando sólo aquellos que comparten los mismos focos geográficos con la consulta que los generó.

El enfoque más simple de *filtrado* es el propuesto por [39], donde su procedimiento es: cuando un documento llega al módulo de filtrado, se identifican todos sus términos

[7]El *rank* es un valor numérico que representa el grado de relevancia que tiene cada documento para la consulta dada.

geográficos, estos son buscados dentro de un catálogo geográfico y si existe más de una definición para dicho término todas son consideradas. Para que un documento pase el filtro, éste debe contener al menos un término geográfico que ya sea coincida con los que aparecen en la consulta o que tenga un parentesco (término padre; e.g., *Paris-Francia*) con los términos geográficos de la consulta. Una estrategia muy similar es presentada en [71] con la diferencia de que se aplica un proceso de desambiguación de los términos geográficos.

Estrategias de *filtrado* más complejas son presentadas en [25, 74, 34], donde el filtrado es realizado a través de varias heurísticas. Generalmente lo primero que se hace es identificar el "tipo" al que corresponde cada término geográfico identificado en los documentos recuperados, siendo estos: *Continente, País, Ciudad, o Lugar*. Este mismo proceso de asignación de tipos es aplicado a la consulta, con el agregado de que se identifica y clasifica a la relación espacial. Así entonces, supongamos que la relación espacial es *"al norte de"* y si el "tipo" asignado al término geográfico es *País*, un ejemplo de la heurística a seguir es sacar las latitudes y longitudes máximas y mínimas que abarcan a dicho país, con esto se calcula un punto medio (en términos de coordenadas geográficas), posteriormente con este punto medio se obtienen los nombres de los lugares que estén contenidos hacia el norte de dicho punto. De esta forma, si un documento no contiene explícitamente el término geográfico de la consulta, pero si alguno de los que se determinó están *"al norte de"* entonces el documento se deja pasar por el filtro[8].

Entre las ventajas que ofrecen las técnicas de filtrado están el hecho de que permiten ir eliminando documentos irrelevantes, sin embargo, estos trabajos dependen en gran medida de la correcta asignación de los focos geográficos y la adecuada interpretación de las relaciones espaciales. Hasta la fecha, la identificación e interpretación automática de las relaciones espaciales sigue siendo un problema abierto en GIR, los trabajos que se describieron cuentan con un listado de relaciones espaciales así como con un conjunto de reglas definidas manualmente para la interpretación de éstas, hecho que no permite utilizar estos sistemas con consultas geográficas más complejas.

Finalmente, técnicas de re-ordenamiento más complejas empleadas por algunos trabajos [55, 58, 16, 67, 5, 66, 62, 47, 91, 63, 70, 15, 69] consisten en definir una función de ordenamiento que básicamente es una combinación lineal de uno o varios valores de similitud, los cuales son asignados a un conjunto de documentos empleando

[8]Para conocer en detalle las heurísticas empleadas refierase a [34]

distintas formas de representación para éstos. En general, la función de *rank* que estos trabajos emplean tiene la siguiente forma:

$$Ranking(q,d) = \lambda(Sim_{tem}(q,d)) + (1-\lambda)(Sim_{geo}(q,d)); \tag{3.2.1}$$

donde $Sim_{tem}(q,d)$ es la similitud temática que existe entre la consulta $q$ y el documento $d$. Generalmente esta similitud es calculada por medio del modelo espacio vectorial (Vea sección 2.1.1), mientras que la similitud geográfica $Sim_{geo}(q,d)$ entre la consulta y el documento puede ser calculada de distintas formas. Por último, el factor lambda ($\lambda$) es solamente un factor de pesado encargado de dar mayor o menor importancia a cada uno de los términos de esta función de ordenamiento.

Una característica común de estos trabajos es la necesidad de contar con focos geográficos en documentos y en la consulta. Como ya hemos mencionado antes, la tarea de asignación de focos geográficos no es una tarea trivial además de que depende del uso de variados y complejos recursos geográficos. Sin embargo, dejando de lado el problema de asignar focos geográficos, y suponiendo que tanto las consultas como los documentos de la colección tienen correctamente asignados sus focos geográficos, los siguientes puntos muestran las medidas geográficas que estos diferentes trabajos han empleado en la función de ordenamiento (Vea fórmula 4.3.1).

- **Similitudes geométricas**: Esta estrategia implica el uso de *Rectángulos Abarcadores Mínimos* o en Inglés *Minimal Bounding Rectangles (MBRs)* para hacer una aproximación de los polígonos de los lugares mencionados en la consulta contra aquellos que son mencionados en los documentos recuperados [55, 58, 69]. La figura 3.2 muestra un ejemplo de un MBR.

  Una vez que se tienen los polígonos de las entidades geográficas mencionadas, o en su defecto la del foco geográfico, se puede determinar si ambos polígonos se traslapan total o parcialmente por medio de emplear las coordenadas de los MBRs. La principal desventaja de emplear MBRs es que tienden a sobre-estimar las áreas de interés, a representar deficientemente áreas irregulares o áreas que involucran múltiples regiones resultando en cálculos poco precisos.

- **Distancia topológica a través de relaciones jerárquicas:** Esta medida es determinada a través de una ontología geográfica, donde la relación *parte de* puede ser utilizada para inferir cierta similitud [5, 16, 15, 66, 47, 69]. Por

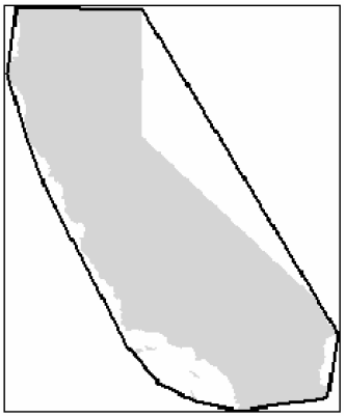

**Figura 3.2: Ejemplo del polígono abarcador mínimo del estado de California E.U**

ejemplo, sabemos que Alicante es parte de España el cual a su vez es parte de Europa, así entonces, Alicante debería ser más similar a España que a Europa. La figura 3.3 muestra un ejemplo de una ontología geográfica, donde se puede observar la jerarquía de los elementos geográficos.

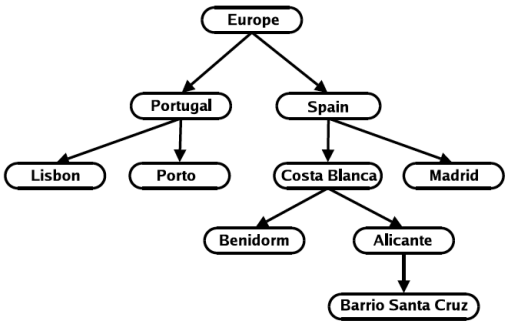


**Figura 3.3: Ejemplo de una ontología geográfica**

- **Distancias Espaciales:** Para el cálculo de distancias espaciales se han definido diferentes heurísticas que tratan de contextualizar la noción de cercanía dependiendo de los focos geográficos en cuestión [5, 16, 15, 66, 47, 70]. Sin embargo, en general para hacer el cálculo de distancias espaciales se emplean las coordenadas geográficas de los focos geográficos las cuales son obtenidas de uno o varios catálogos geográficos.

- **Población compartida:** Esta medida considera que cuando dos regiones están

conectadas a través de una relación *parte-de*, la fracción de población del área más general es también asignada al área más particular, y puede ser utilizada para calcular una medida de similitud [16, 15, 66, 47]. En otras palabras, esta medida es una forma de determinar la importancia relativa de una región dentro de otra y es vista como una forma alternativa para determinar el nivel de traslape entre dos regiones.

- **Adyacencia dentro de una ontología:** Esta medida se basa en el hecho de que dos lugares adyacentes dentro de una ontología son más similares que dos lugares que no lo sean. Empleando la relación de *adyacencia* de la ontología es posible determinar un valor de similitud entre dos lugares [5, 16, 15, 66]. Así entonces, tomando como ejemplo la ontología mostrada en la figura 3.3, el término "Madrid" tendrá un valor de adyacencia (similitud) alto con el término "Spain", mientras que con el término "Costa Blanca" este valor será mucho menor.

Una vez calculadas estas diferentes medidas, estas son empleadas ya sea de manera individual en la formula 4.3.1 o combinadas de forma que al final se tenga un único valor para el término $Sim_{geo}(q, d)$. Sin embargo, estas medidas dependen en su mayoría de haber asignado correctamente los focos geográficos o en su defecto de haber desambiguado correctamente las entidades geográficas. Agregado a este problema, está presente el problema de la cobertura de los recursos geográficos, pues los recursos disponibles abarcan subconjuntos del total de nombres de lugares existentes en el mundo [11], como ejemplo de esto, la figura 3.4 muestra la cobertura que tiene la base de datos de Wikipedia[9]

El uso de estas medidas de similitud geográfica, ya sea de manera individual o combinadas, no han mostrado ser de gran utilidad, pues en varios de los trabajos que proponen y emplean estas medidas en su función de ordenamiento, el asignar un $\lambda$ menor, es decir, darle mayor importancia al término geográfico $Sim_{geo}(q, d)$ (formula 4.3.1) resulta en la degradación del orden de los documentos recuperados. Por si esto fuera poco, no hay un claro estudio sobre cual de estas medidas es realmente la que aporta mayor información en el proceso del re-ordenamiento.

Nuestra hipótesis principal sobre el porqué estas estrategias de ordenamiento no han mostrado ser efectivas, es debido a la falta de información en la consulta. Así en-

[9]Wikipedia World Database

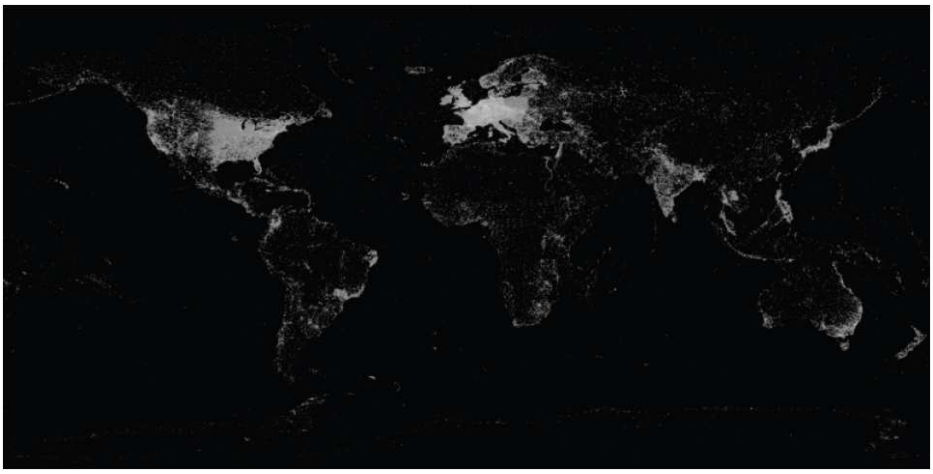

**Figura 3.4: Cobertura de la Wikipedia World database (Puntos iluminados corresponden a los lugares sobre los que se tiene registro en la base de datos)**

tonces, la idea central de este trabajo es que por medio de utilizar *documentos ejemplo* como fuente de información en el proceso del re-ordenamiento, es posible obtener una aproximación en forma más explícita de la información implícita contenida en las consultas geográficas.

## 3.3. Discusión

En este capítulo se revisaron trabajos que se han enfocado en el desarrollo y mejoramiento de los sistemas de recuperación de información geográfica. Algunos aspectos importantes a resaltar son la gran cantidad de sub-problemas que son necesarios resolver para poder contar con un sistema GIR robusto. En la figura 3.1 se mostró nuestra visión de los componentes mínimos que un sistema GIR debe contener para poder resolver efectivamente las necesidades de información de un usuario en el contexto de búsquedas geográficas.

En la Tabla 3.1 se muestra de manera muy general las diferentes configuraciones empleadas por los grupos de investigación más representativos dentro del foro del GeoCLEF[10]. Esta tabla tiene la finalidad de proporcionar al lector de una visión de la variedad de configuraciones empleadas por algunos de los trabajos revisados en este capítulo, así como de su desempeño obtenido dentro de la tarea de recuperación

[10]La descripción de la tarea del GeoCLEF es dada con mayor detalle en la sección 5.2

de información geográfica.

En el primer módulo de nuestra arquitectura, el *análisis de la consulta* (Vea Figura 3.1), los principales problemas a resolver son identificar adecuadamente las partes de la consulta, i.e., partes *temáticas* y *geográficas*. Un sistema ideal, posterior a esta identificación, debería ser capaz de hacer una expansión adecuada de los términos de la consulta. Aquí el problema principal es la correcta identificación y desambigüación de los términos geográficos. No obstante, una característica de los trabajos mostrados en la tabla 3.1 es la preferencia por la expansión de los términos *temáticos* de la consulta, siendo estos mismos los que en su mayoría obtienen mejores resultados.

Agregado a este problema se encuentra el de cómo tratar la *relación espacial*. Como se vio en los primeros capítulos, la relación espacial juega un papel importante en la correcta interpretación de la necesidad de información del usuario. A la fecha, algunos tipos de *relaciones espaciales* son procesadas por medio de reglas muy específicas, sin embargo, los grupos de investigación que hacen esto no obtienen mejoras significativas en sus resultados (Vea Tabla 3.1).

Para el caso del segundo módulo, la *máquina de recuperación de información*, los principales problemas a resolver son el cómo generar índices geográficos, de tal forma que sea posible recuperar documentos relacionados geográficamente entre si sin la necesidad de que éstos documentos hagan una mención explícita de dicha relación. Para alcanzar este fin es inevitable el uso de uno o varios recursos geográficos con los cuales sea posible conocer estas relaciones además de que es necesario contar con procesos eficientes de desambigüación de términos geográficos los cuales permitan "geo-etiquetar" adecuadamente la colección de documentos. Existen algunos grupos de investigación que han tratado de resolver el problema GIR aplicando estas ideas, sin embargo, como se puede observar en la tabla 3.1, este tipo de configuración no les ha permitido obtener mejoras significativas comparado contra aquellos grupos que no lo hacen.

Finalmente, en el módulo del ordenamiento por relevancia, el principal problema es otorgar a los documentos el *orden* adecuado a los requerimientos de información (*temáticos* y *geográficos*) del usuario. Como se vio en la sección 3.2.3 diferentes estrategias han sido propuestas para realizar este ordenamiento. Desde la aplicación de operadores lógicos a diferentes conjuntos de listas de documentos hasta definición de complejas fórmulas para determinar la cercanía temática y geográfica de los documentos recuperados. En la tabla 3.1 todos los trabajos que hacen re-ordenamiento, a

excepción de XLDB-06 [66] y TALP [25], lo hacen por medio de operaciones sencillas de fusión de listas y sin embargo con sus resultados obtenidos lograron colocarse en la lista de los sistemas mejor evaluados.

En general, las conclusiones a las que llegamos a partir del estudio realizado en este capítulo son:*i)* es conveniente considerar tanto lo *temático* como lo *geográfico* en el proceso general de recuperación, pues son partes complementarias; *ii)* independientemente de la configuración empleada por las máquinas de IR, éstas son capaces de recuperar documentos relevantes, el problema es que no les están otorgando un orden adecuado. Por ultimo, *iii)* los términos aislados de las consultas son insuficientes para realizar el ordenamiento de los documentos.

| **Nombre del Sistema** | **Expansión** *Temática* | *Geográfica* | **Múltiples índices** | **Aplica un re-ordenamiento** | **Año consultas GeoCLEF** | **MAP** |
|---|---|---|---|---|---|---|
| **BERKELEY [56]** | ✓ | | | | **2005** | **0.3936** |
| CSUSM-05 [37] | ✓ | | | | 2005 | 0.3613 |
| ALICANTE-05 [23] | ✓ | | | | 2005 | 0.3495 |
| CHESHIRE-05 [52] | | ✓ | ✓ | ✓ | 2005 | 0.2924 |
| MIRACLE [50] | | ✓ | | ✓ | 2005 | 0.2653 |
| **XLDB-06 [66]** | ✓ | | ✓ | ✓ | **2006** | **0.3034** |
| ALICANTE-06† [82] | | ✓ | | ✓ | 2006 | 0.2723 |
| CSUSM-06 [33] | ✓ | | | | 2006 | 0.2637 |
| UNSW† [44] | | | ✓ | ✓ | 2006 | 0.2622 |
| SINAI† [30] | | ✓ | | | 2006 | 0.2611 |
| **TALP [25]** | ✓ | | ✓ | ✓ | **2007** | **0.2850** |
| CHESHIRE-07 [53] | ✓ | | | | 2007 | 0.2642 |
| UPV [13] | | | | ✓ | 2007 | 0.2636 |
| CLCG [4] | | ✓ | | | 2007 | 0.2515 |
| CSUSM-07 [38] | ✓ | | | | 2007 | 0.2132 |
| **DFKI [90]** | | ✓ | | | **2008** | **0.3037** |
| ALIVALE [9] | ✓ | | | ✓ | 2008 | 0.2864 |
| JAEN† [75] | | ✓ | | | 2008 | 0.2841 |
| XLDB-08 [17] | | ✓ | ✓ | | 2008 | 0.2755 |
| CHESHIRE-08 [54] | ✓ | | | ✓ | 2008 | 0.2685 |

[11] † Grupos que definen reglas para la identificación y el procesamiento de la relación espacial.

**Tabla 3.1: Descripción general de los trabajos más representativos del foro de evaluación GeoCLEF**

## Capítulo 4

# Ordenamiento Basado en Ejemplos

En este capítulo se describe en detalle el sistema GIR propuesto. Como ya se ha mencionado antes, nuestro trabajo de investigación se enfocó principalmente en el módulo de ordenamiento por relevancia. La principal característica de nuestro sistema es el uso de *documentos ejemplo* para el proceso de re-ordenamiento. Adicionalmente, nuestro trabajo se diferencia de los hasta ahora existentes en el hecho de que empleamos la información obtenida a través de un proceso de retroalimentación de relevancia para hacer el ordenamiento y no para realizar un segundo proceso de recuperación, evitando así las desventajas que traen consigo las técnicas de reformulación de consultas vía retroalimentación de relevancia.

La revisión realizada a los trabajos más sobresalientes del área GIR (Véase capítulo 3) nos permitió llegar a la conclusión de que sistemas tradicionales de IR son capaces de recuperar documentos relevantes a consultas geográficas, sin embargo, estos sistemas no son capaces de generar un orden pertinente de los documentos recuperados, resultando esto en un desempeño deficiente del sistema GIR. En otras palabras, generalmente se tienen niveles aceptables de *recuerdo* pero la *precisión* del sistema es baja. Nuestra hipótesis es que este comportamiento se da por dos principales razones: *i)* debido a la falta de información en las consultas geográficas; *ii)* la presencia de requerimientos de información en forma implícita dentro de las consultas.

En la primera sección de este capítulo se muestra de manera general la arquitectura propuesta. Posteriormente se describen las formas que se emplearon para manejar y mezclar la información proporcionada por los *documentos ejemplo* dentro del módulo de re-ordenamiento. Finalmente, en la tercera sección se describen las formas de representación empleadas para poder medir similitudes entre *documentos ejemplo* y el conjunto de documentos recuperados.

## 4.1. Arquitectura Propuesta

El método propuesto consiste de dos etapas principales: *i) etapa de recuperación* y *ii) etapa de re-ordenamiento*. El objetivo de la primera etapa es, precisamente como su nombre lo indica, recuperar por medio de la **Máquina de IR** tanto como sea posible documentos relevantes a consultas geográficas, mientras que el objetivo de la segunda etapa es mejorar el orden asignado a los documentos recuperados (i.e., un **Re-ordenamiento**) por medio de emplear la información contenida en el conjunto de *documentos ejemplo*, los cuales son obtenidos a partir del módulo de **Retroalimentación de Relevancia**. Un diagrama general de la arquitectura propuesta se muestra en la figura 4.1.

Así entonces, dada una consulta geográfica, el sistema GIR recupera de la colección de documentos un conjunto de elementos los cuales son entregados en un determinado orden definido por algún criterio de recuperación (Vea sección 2.1). De esta lista de documentos recuperados, algunos elementos (presumiblemente relevantes) son seleccionados a través de una estrategia de retroalimentación de relevancia (*simulada* o *ciega*), a estos los denominamos como *documentos ejemplo* . Los *documentos ejemplo* junto con los documentos recuperados son enviados al módulo de re-ordenamiento, donde básicamente la idea es re-ordenar los documentos recuperados de acuerdo al grado de similitud o cercanía que tengan con los *documentos ejemplo*.

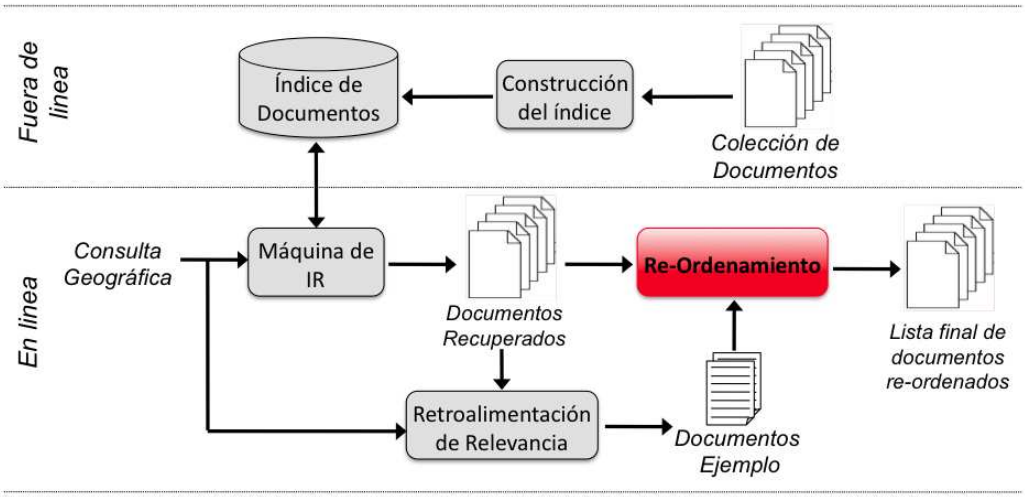


**Figura 4.1: Diagrama de bloques de la arquitectura propuesta**

Dentro de esta arquitectura nos enfrentamos al problema de cómo manejar la información de los *documentos ejemplo*, i.e., de qué forma los vamos a fusionar de tal manera que sea posible medir la similitud de los documentos recuperados contra éstos. En la sección 4.2 se presentan tres diferentes formas que se propusieron para el manejo de la información contenida en los *documentos ejemplo*.

Posteriormente, una vez definida la forma en que se manejaría la información contenida en los *documentos ejemplo* nos enfrentamos a la incógnita de cómo medir la similitud de los documentos recuperados contra los *documentos ejemplo*. En la sección 4.3 se describen las diferentes formas en que se representó tanto a los documentos recuperados como a los *documentos ejemplo* para poder medir la similitud entre éstos. Así entonces, las siguientes secciones describen el trabajo realizado dentro del módulo de re-ordenamiento de la figura 4.1.

## 4.2. Manejando los Documentos Ejemplo

La idea de manejar *documentos ejemplo* en el proceso de recuperación no es nueva, este esquema es ampliamente usado en el área de recuperación de imágenes [26]. Esta idea surge debido a la complejidad que involucra para el usuario el redactar una consulta efectiva que le permita recuperar imágenes que satisfagan su necesidad de información. De esta forma, si el usuario en lugar de proporcionar una consulta escrita proporciona un ejemplo del tipo de imágenes que quiere recuperar, el sistema de IR puede hacer una mejor aproximación de la información de los ejemplos que de una consulta escrita, permitiéndole producir una salida de mayor calidad.

De aquí nuestra idea principal, pues dado que consultas geográficas tienden a contener requerimientos de información en forma implícita, suponemos que el uso de *documentos ejemplos* ayudará en el proceso general de recuperación de información, i.e., ayudarán a obtener una aproximación en forma explícita de la información implícita contenida en las consultas. Nótese que a diferencia de lo que se hace en recuperación de imágenes, nuestro trabajo propone usar ejemplos en el proceso del re-ordenamiento por relevancia y no en el proceso de recuperación. En este sentido, nuestra idea ayudará de manera particular a mejorar la calidad de la salida (i.e., *precisión*) de los sistemas actuales de recuperación de información en el contexto de búsquedas geográficas.

Además, a diferencia de IR tradicional, la información obtenida a través del proce-

so de retroalimentación de relevancia la usaremos para el proceso del re-ordenamiento y no para reformular consultas. De esta forma evitamos las desventajas que trae realizar un segundo proceso de recuperación con la reformulación de consultas vía retroalimentación de relevancia, entre las cuales podemos mencionar la alta sensibilidad de estas estrategias a la calidad de las palabras agregadas y el hecho de que el resultado de la recuperación puede verse desviado de la intención de búsqueda inicial si demasiadas palabras son agregadas. En pocas palabras, el efectivo desempeño de las estrategias de reformulación de consultas vía retroalimentación de relevancia dependen en gran medida de la calidad del conjunto inicial de documentos recuperados [93].

Una vez definida la arquitectura básica de nuestro sistema GIR (figura 4.1) nuestro primer problema a resolver fue el definir la forma de cómo manejar la información contenida en los *documentos ejemplo*. Para esto, proponemos tres formas distintas de manejar la información de los documentos ejemplos, las cuales son descritas a continuación.

### 4.2.1. Documento Virtual

Esta idea surge de estudiar los sistemas de recuperación de imágenes, donde estos sistemas reciben como consulta una imagen de lo que el usuario está buscando. Generalmente esta imagen es conservada en su totalidad, pues el sistema supone que ésta contendrá información explícita que ayudará a recuperar imágenes altamente relacionadas con la necesidad de información del usuario. Entonces, la idea de construir un "documento virtual" es con la finalidad de conservar la mayor información posible de los *documentos ejemplo*. Por medio de este "documento virtual" se espera obtener una aproximación explícita de las necesidades de información del usuario y posteriormente generar un orden más pertinente de la lista de documentos recuperados.

La idea básica de esta forma de manejar la información contenida en los *documentos ejemplo* consiste en, dados $n$ documentos ejemplo, llamaremos "documento virtual" al resultado de la concatenación de estos $n$ ejemplos seleccionados en el módulo de retroalimentación. En otras palabras, la salida de este proceso de construcción de un "documento virtual" será un "único" documento que contiene en su interior toda la información de los ejemplos seleccionados.

De manera más formal tenemos, dado un conjunto $R = \{r_1, r_2, ..., r_m\}$ de documentos recuperados, se seleccionan de este conjunto un total de $n$ *documentos ejemplo*,

los cuales están representados por el conjunto $E = \{e_1, e_2, ..., e_n\}$ donde se cumple que $E \subset R$ y además $n \ll m$. Entonces, definimos al documento virtual $v$ como:

$$v = e_1 \cup e_2 \cup ... \cup e_n \tag{4.2.1}$$

Si consideráramos que estamos representando[1] a los documentos por medio del modelo espacio vectorial (VSM), cada documento ejemplo está representado por un vector $\vec{e_i}$, así entonces otra forma de ver al documento virtual $v$ es como el resultado de una suma de vectores (fórmula 4.2.2).

$$\vec{v} = \vec{e_1} + \vec{e_2} + ... + \vec{e_n} \tag{4.2.2}$$

Finalmente, dado que podemos contar con las representaciones vectoriales de los documentos recuperados ($R$), es posible compararlos contra el documento virtual $\vec{v}$, calcular un valor de similitud a cada documento $\vec{r_i}$ y con él re-ordenar la lista de documentos recuperados.

### 4.2.2. Resumen Multi-documento

Dado que los documentos ejemplo son seleccionados de un conjunto de documentos recuperados ($R$), los cuales fueron obtenidos por una máquina de IR tras responder a una consulta geográfica, se supone que todos estos documentos guardan una relación temática. Siendo esto cierto, el conjunto de documentos ejemplo ($E$) contiene información redundante así como también información diferente que puede ser importante.

Es bajo estos supuestos que surge la idea de aplicar una estrategia de generación de resúmenes de múltiples documentos para manejar la información contenida en los ejemplos seleccionados. La salida de este proceso es un "documento corto" en el cual las redundancias de información presentes en el conjunto $E$ serán resueltas, además de que también las diferencias importantes serán consideradas para la construcción de dicho resumen multi-documento.

El propósito de hacer un resumen es eliminar la información contenida en los *documentos ejemplo* que no aporte elementos suficientes para hacer un ordenamiento

[1] En la sección 4.3 se exponen los diferentes tipos de información empleada para la representación de los documentos así como la forma en que se midió la similitud entre estos.

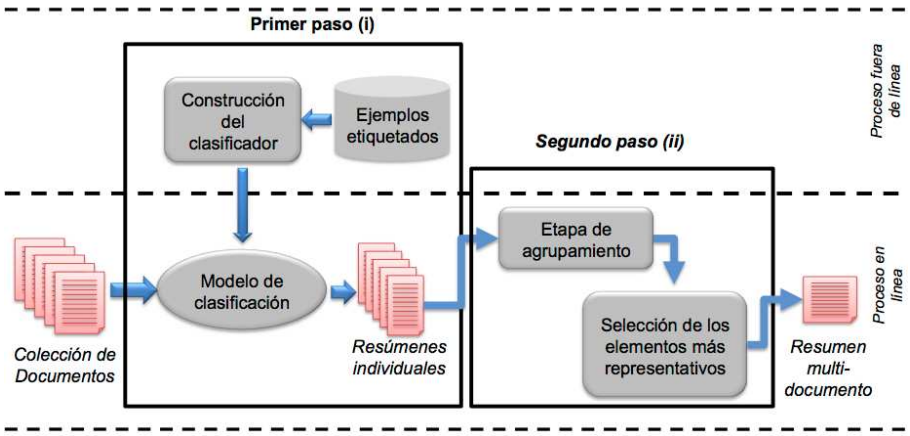


**Figura 4.2: Arquitectura del método de generación de resúmenes de múltiples documentos basado en la detección de oraciones relevantes locales**

adecuado. Tras haber realizado un análisis de la información contenida en los *documentos ejemplo*, fue posible notar que existían gran cantidad de términos geográficos que no tenían relación directa con el foco geográfico de la consulta dada al sistema. Entonces, la finalidad del resumen es concentrar la información similar de los *documentos ejemplo* y dejar fuera la información contradictoria o poco relevante, en particular la información geográfica, pues se espera que en el resumen se conserven sólo aquellas referencias geográficas relevantes.

Para la realización de los resúmenes de múltiples documentos empleamos una técnica basada en la detección de oraciones localmente relevantes [86]. La figura 4.2 muestra un diagrama de bloques del método empleado.

Este método está dividido en dos grandes etapas, *i)* primero, un proceso supervisado se encarga de tratar a cada documento de la colección de documentos a resumir de manera individual, en este paso se detectan las oraciones relevantes de cada uno de éstos, i.e., las *oraciones localmente relevantes*. La salida de este primer paso es un conjunto de resúmenes individuales e cada uno de los documentos de entrada al sistema. Posteriormente, *ii)* un proceso no supervisado se encarga de localizar todos los temas contenidos dentro de los resúmenes individuales generados en el paso anterior, de tal forma que una vez hecho esto, es posible seleccionar los elementos más representativos de cada tema para así crear el resumen multi-documento final.

- **Primer paso: extracción de las oraciones localmente relevantes.** El módulo principal del sistema de generación de resúmenes multi-documento

está contenido en este primer paso. Este módulo se concentra en la creación de resúmenes individuales (i.e., resúmenes de un sólo documento) por medio de la selección de las oraciones relevantes contenidas en cada uno de los documentos de entrada.

La técnica empleada por este módulo consiste en una técnica supervisada para la generación de resúmenes de un sólo documento, donde los documentos son representados a través de secuencias de palabras ($n$-gramas). La idea de emplear palabras como atributos tiene como finalidad permitirle al sistema de generación de resúmenes ser más flexible y menos dependiente del dominio y del lenguaje [85].

En particular, se emplearon $n$-gramas (secuencias de $n$ palabras consecutivas) como atributos para representar a las oraciones. De esta forma, en el modelo de clasificación, cada oración de cada documento es representada por un vector de atributos que contiene un valor booleano por cada $n$-grama que ocurre en el conjunto de entrenamiento. Para los experimentos realizados se consideraron secuencias de $n$-gramas con $1 \leq n \leq 3$.

Como clasificador se empleó la estrategia naïve Bayes, donde se calcula para cada oración $s$ su probabilidad de pertenecer al resumen $S$ dados $k$ atributos los cuales están representados por $F_j; j = 1...k$. Esta probabilidad puede ser expresada empleando la regla de Bayes de la siguiente forma:

$$P(s \in S|F_1, F_2, ..., F_k) = \frac{P(F_1, F_2, ..., F_k|s \in S)P(s \in S)}{P(F_1, F_2, ..., F_k)} \tag{4.2.3}$$

Y suponiendo que los atributos son independientes tenemos:

$$P(s \in S|F_1, F_2, ..., F_k) = \frac{\prod_{j=1}^{k} P(F_j|s \in S)P(s \in S)}{\prod_{j=1}^{k} P(F_j)} \tag{4.2.4}$$

donde $P(s \in S)$ es un valor constante y $P(F_j|s \in S)$ y $P(F_j)$ pueden ser calculadas directamente del conjunto de entrenamiento.

- **Segundo paso: identificación de temas.** Este segundo paso involucra aplicar las técnicas tradicionalmente empleadas para resolver el problema de generación de resúmenes de múltiples documentos. Para este proceso se empleó el *algoritmo*

*de agrupamiento estrella* [86], entre las ventajas que tiene este algoritmo son que: *i)* induce de manera natural el número de grupos, *ii)* identifica la estructura jerárquica de los temas dentro del espacio de documentos.

El algoritmo de agrupamiento estrella está basado en una cobertura ávida[2] de un grafo de similitudes umbralizado $G_\sigma$ generando como salida grupos de subgrafos en forma de estrella. El elemento central de cada estrella es llamado *centro de la estrella*, mientras que los elementos que lo rodean son denominados *satélites*.

El grupo de estrellas generadas se dice que son correctas si:(1) cada centro de estrella no es adyacente a ningún otro centro de estrella, y (2) si cada satélite es adyacente al menos a un centro de estrella con igual o mayor grado.

El procedimiento que se siguió para los experimentos fue: calcular similitudes entre las oraciones de los documentos de entrada empleando la medida DICE (Véase formula 2.1.3 ) dando como resultado una matriz $M$ de similitudes. La construcción del grafo umbralizado $G_\sigma$ se hizo definiendo el umbral $\sigma$ por medio de $\sigma = \bar{x} + \delta$, donde $\bar{x}$ es el valor de la media estadística y $\delta$ es la desviación estándar, ambos calculados a partir de la matriz de similitud $M$. Al definir el umbral $\sigma$ de esta manera, nuestra intención es hacer el método adaptable y adecuado a la naturaleza de los documentos, además de evitar la intervención del usuario en el proceso de definir un umbral apropiado.

La metodología que se siguió para la construcción del resumen multi-documento final fue: desde la estrella más grande a la más pequeña se tomó la oración colocada como "centro de estrella". Si tras esto, el tamaño predefinido del resumen no ha sido alcanzado, se vuelven a revisar las estrellas, pero esta vez se toman las oraciones "satélite" (satélites que no hayan sido considerados antes) con mayor grado de similitud al "centro de estrella". Este paso es repetido hasta alcanzar el tamaño del resumen deseado.

Finalmente, el resumen generado $s$ puede ser llevado a un tipo de representación vectorial o alguna otra de las expuestas en la sección 4.3, y de manera similar a lo que hacemos con el *documento virtual* expuesto en la sección anterior, $s$ es comparado contra el conjunto de documentos recuperados $R$ de tal forma que el valor de similitud

[2]El nombre de algoritmos ávidos, también conocidos como voraces (del término inglés "greedy") se debe a su comportamiento: en cada etapa el algoritmo "toma lo que puede o la mejor solución en ese instante" sin analizar consecuencias, es decir, son glotones por naturaleza.

obtenido tras esta comparación servirá como criterio para generar el re-ordenamiento de la lista de documentos recuperados.

### 4.2.3. Campos de Markov

La idea de manejar campos aleatorios de Markov surge al notar que en nuestras dos propuestas anteriores de cómo manejar los *documentos ejemplo* (i.e., un documento virtual y un resumen multi-documento) no se está tomando en cuenta información que pudiera aportar elementos importantes para el proceso de re-ordenamiento. Así entonces, la finalidad de este punto es integrar mediante el uso de un MRF, el orden original obtenido por la máquina de IR, la similitud entre los documentos recuperados y los *documentos ejemplo*, todo con el propósito de generar un orden más apropiado para la lista de documentos recuperados.

El MRF debe separar los documentos relevantes de los que no lo son mediante el uso de una función de energía, la cual debe representar la similitud entre los documentos de la misma clase (relevantes y no-relevantes) haciendo uso de algún conjunto de características (e.g., palabras). La función de energía también debe representar la información *a priori* de los documentos, representada como la similitud con la consulta o el orden original generado por la máquina de IR. El uso de los *documentos ejemplo* permite incluir la intención de búsqueda en la función de energía.

Los MRFs modelan procesos en los que las variables aleatorias, además de tener una probabilidad asociada también se ven influenciadas por la probabilidad conjunta de sus variables vecinas (Véase sección 2.3). Dadas las probabilidades individuales de las variables y las probabilidades de sus vecinos, lo que se quiere encontrar es la configuración más probable.

Para construir el MRF se representa a cada documento como una variable aleatoria de dos posibles valores *relevante* y *no-relevante*. Para comparar la relación entre todos los documentos se definió un esquema de vecindad en el que cada variable tiene como vecinas al resto de las variables.

La función de energía empleada toma en cuenta dos factores: la diferencia entre los documentos recuperados (*diferencia interna*); y la información externa, obtenida de el orden inicialmente asignado a los documentos recuperados y la diferencia de cada documento con los *documentos ejemplo* (*diferencia externa*) [19, 84]. La diferencia interna corresponde al potencial de interacción mientras que la diferencia externa al

potencial de asociación.

La función de energía se define de la siguiente forma:

$$U(F) = \varphi V_c(F) + (1 - \varphi)V_a(F) \tag{4.2.5}$$

donde $V_c$ es el potencial de interacción e indica la diferencia entre la variable aleatoria $F$ y sus vecinas, lo cual representa el apoyo que dan los vecinos de una variable para que cambie o mantenga su valor. $V_a$ es el potencial de asociación y representa la influencia de la información externa representada por el ordenamiento original de la máquina de IR y la diferencia de cada documento con los *documentos ejemplo*. Finalmente, $\varphi$ es un factor que le da más relevancia al potencial de interacción o al potencial de asociación con valores entre 0 y 1.

El potencial $V_c$ se compone de dos elementos. El primero representa el apoyo que le dan las variables vecinas a la variable $F_i$ con el mismo valor; mientras que el segundo representa el rechazo que le dan las variables vecinas a $F_i$ con valor distinto. Ambos elementos hacen uso de una medida de diferencia ($dif$), que puede emplear distintos atributos para representar a los documentos (Véase sección 4.3) y sobre estos calcular esta diferencia. Así entonces, el potencial $V_c$ se define como:

$$V_c(F_i) = \begin{cases} \frac{\sum_j^M dif(F_i,F_j)}{M} + (1 - \frac{\sum_j^N dif(F_i,F_j)}{N}), & \text{si } F_i \text{ es no relevante} \\ \frac{\sum_j^N dif(F_i,F_j)}{N} + (1 - \frac{\sum_j^M dif(F_i,F_j)}{M}), & \text{si } F_i \text{ es relevante} \end{cases} \tag{4.2.6}$$

donde $N$ representa el número de variables vecinas de $F_i$ con valor relevante, y $M$ el número de variables vecinas de $F_i$ con valor no-relevante.

El potencial de asociación $V_a$ también emplea dos elementos para su cálculo. El primero representa la diferencia entre el documento $F_i$ y los *documentos ejemplo*; dado que los *documentos ejemplo* pueden ser uno o más, para poder determinar esta diferencia los *documentos ejemplo* fueron llevados a su forma de "documento virtual" ($v$). Por otro lado, el segundo elemento fija un valor numérico real a la posición del documento $F_i$ en la lista original. Así entonces, $V_a$ se define de la siguiente forma:

$$V_a(F_i) = \begin{cases} (1 - dif(F_i, v)) \times \theta(posinv(F_i)), & \text{si } F_i \text{ es } \textit{no relevante} \\ dif(F_i, v) \times \theta(pos(F_i)), & \text{si } F_i \text{ es } \textit{relevante} \end{cases} \tag{4.2.7}$$

La función $\theta(pos(F_i))$ mapea la posición $pos(F_i)$ o la posición inversa $posinv(F_i)$ de $F_i$ en la lista de documentos recuperados a un valor numérico real para apoyar o castigar al documento de acuerdo a su posición [19].

Una vez descritos los potenciales, la función de energía empleada se define como:

$$U(F_i) = \begin{cases} \varphi(\frac{\sum_j^M dif(F_i,F_j)}{M} + (1 - \frac{\sum_j^N dif(F_i,F_j)}{N})) + (1-\varphi)[1 - dif(F_i, v) \times \theta(posinv(F_i))], \\ \text{si } F_i \text{ es } \textit{no relevante} \\ \\ \varphi(\frac{\sum_j^N dif(F_i,F_j)}{N} + (1 - \frac{\sum_j^M dif(F_i,F_j)}{M})) + (1-\varphi)[dif(F_i, v) \times \theta(pos(F_i))], \\ \text{si } F_i \text{ es } \textit{relevante} \end{cases} \tag{4.2.8}$$

La configuración inicial del MRF es obtenida a través de los elementos seleccionados por la estrategia de retroalimentación de relevancia, i.e., los *documentos ejemplo*. El conjunto de *documentos ejemplo* son inicializados como *relevantes*, mientras que el resto de los documentos en la lista son marcados como *no-relevantes*. Posteriormente, el MRF se iteró a través del algoritmo ICM, el cual prefiere el valor de la variable con el que se obtenga menor energía. Al final de este proceso de optimización, cada variable (documento) tiene un valor de relevante o no relevante. Con base en estos valores una nueva lista re-ordenada es producida.

Finalmente es conveniente mencionar que la medida de diferencia ($dif(F_i, F_j)$) empleada para los experimentos se define como:

$$dif(F_i, F_j) = 1 - sim(F_i, F_j) \tag{4.2.9}$$

donde $sim(F_i, F_j)$ indica el valor de similitud entre los documentos $F_i$ y $F_j$. Es importante tener en cuenta que este valor de similitud variará dependiendo de la forma de representación empleada para los documentos. La siguiente sección describe los diferentes tipos de representación empleados para la realización de los experimentos.

## 4.3. Representando los Documentos Ejemplo

En la sección anterior hemos explicado las tres diferentes formas que se propusieron para el manejo de la información contenida en los documentos ejemplo. En esta sección describiremos cómo se representó a dicha información, i.e., qué características de los documentos fueron tomadas en cuenta para poder medir similitudes entre éstos y/o

diferencias que sería el caso particular de los MRFs.

**BOW.** En el tipo de representación *Bag of Words* (BOW por sus siglas en Inglés) o Bolsa de Palabras, la información que se empleó para representar tanto a los *documentos ejemplo* como a los documentos recuperados fue por medio de todas las palabras contenidas en ellos. Así entonces, cuando hablemos de que se empleó una forma de representación de bolsa de palabras o **BOW** quiere decir que el proceso que se siguió para medir similitudes entre documentos se hizo por medio de llevar los documentos a su forma vectorial (sección 2.1.1) y aplicarles la fórmula del *coseno* (fórmula 2.1.2) para determinar dicho valor de similitud.

**Geográfica Simple.** La información empleada para esta forma de representación considera únicamente las "entidades geográficas verificadas" contenidas tanto en los *documentos ejemplo* como en los documentos recuperados. Para lograr este tipo de representación fue necesario el uso de dos recursos adicionales, *i)* un identificador de entidades nombradas (NER) y, *ii)* un catálogo geográfico.

El procedimiento que se siguió consistió en: (1) primero etiquetar todos los documentos con sus entidades nombradas, para lo cual empleamos la herramienta NER de la Universidad de Stanford[3]. Entre las etiquetas que asigna esta herramienta está la de "LUGAR", que es la que nos interesa. Posteriormente, (2) la existencia de cada una de las entidades nombradas de tipo "LUGAR" identificadas en el paso anterior, es "verificada" en el catálogo geográfico Geonames[4], si alguna de las entidades no existe en el catálogo geográfico, ésta es eliminada, en caso contrario ésta es conservada como parte del conjunto de palabras *geográficas simples*.

Así entonces, al final los documentos son llevados a su forma vectorial, donde a diferencia de la representación textual, aquí cada componente del vector son únicamente palabras *geográficas simples* y por medio de la fórmula del coseno se determinan similitudes entre los *documentos ejemplo* y los documentos recuperados. También es importante mencionar que en este tipo de representación se emplean valores *tf-idf* calculados a nivel local, i.e., los valores *tf-idf* de las palabras *geográficas simples* se determinan empleando el conjunto de documentos

[3]http://nlp.stanford.edu/software/CRF-NER.shtml

[4]Geonames es una base de datos geográfica que contiene más de 6 millones de entradas correspondientes a nombres geográficos. Es un recurso disponible en: http://www.geonames.org/

$R$ y no toda la colección de documentos.

**Geográfico Expandido.** Este tipo de representación, similar a la anterior, considera además de las palabras "geográficas simples" , los términos *padre* de cada entidad geográfica. Por ejemplo, si el término "Córdoba" aparece dentro de un documento, los dos ancestros (o términos *padre*) directos de éste son "Veracruz" y "México", los cuales son agregados al documento. Este proceso es repetido para cada entidad geográfica de tal forma que al final se tiene un documento geográficamente expandido.

Nótese que para hacer una adecuada expansión es necesario aplicar un proceso de desambiguación de términos geográficos. Siguiendo con el ejemplo anterior, el término "Córdoba" es ambiguo, pues este mismo nombre puede referirse a una provincia Argentina, o a una población en España, entre otros.

Para poder hacer una desambiguación geográfica seguimos las ideas propuestas en [12]. Donde dado $t$, un término geográfico que es ambiguo que puede tomar $k+1$ valores (significados geográficos),i.e., $t_0, t_1, ..., t_k$; junto con un conjunto de $n+1$ términos geográficos del contexto $C$ ($c_i \in C; 0 \leq i \leq n$), la idea es asignar a $t$ el significado $t_k$ que esté más cercano geográficamente a los términos del contexto. El algoritmo para la desambiguación geográfica está descrito a continuación:

1.- Localizar dentro de algún recurso geográfico las coordenadas geográficas para cada $c_i$. Si $c_i$ es un término ambiguo, considerar todos sus posibles valores. Al conjunto de coordenadas obtenido se le denominará $P_c$

2.- Calcular el centroide $\widehat{c}$ del contexto geográfico, donde $\widehat{c} = \frac{(p_1+p_2+...+p_n)}{n}$ tal que $p_i \in P_c$

3.- Quitar del conjunto $P_c$ todos aquellos puntos que estén a más de $2\sigma$ de distancia de $\widehat{c}$, y re-calcular nuevamente $\widehat{c}$ con los puntos restantes $P'_c$. $\sigma$ es la desviación estándar del conjunto de puntos.

4.- Calcular las distancias desde $\widehat{c}$ a cada $t_0, t_1, ..., t_k$

5.- Seleccionar el significado $t_j$ que tenga la distancia mínima desde $\widehat{c}$

Así entonces, una vez hecha la desambiguación, cada término geográfico del documento es expandido por medio de agregar sus dos ancestros inmediatos, los cuales son extraídos del recurso geográfico Geonames. Finalmente, cada *documento ejemplo* geográficamente expandido es llevado a una representación vectorial y la similitud contra el conjunto de documentos recuperados es medida por medio de la fórmula del coseno. De manera similar al caso de la representación *geográfica simple* aquí también los valores *tf-idf* son determinados a nivel local, i.e., sobre el conjunto $R$.

**Coordenadas Geográficas.** Este tipo de representación no involucra el uso de términos geográficos de manera explícita, en su lugar, lo que se empleó para representar a cada documento (ya sean *documentos ejemplo* o documentos recuperados) es un par de coordenadas geográficas (latitud y longitud), las cuales son determinadas por medio de considerar todos los términos geográficos contenidos en los documentos.

El procedimiento para lograr este tipo de representación fue: (1) identificar las entidades geográficas para cada documento, (2) verificar la existencia de dichas entidades en el catálogo geográfico, (3) desambiguar por medio del algoritmo expuesto en el punto anterior los términos geográficos que lo requieran, y (4) con el conjunto de términos geográficos ya verificados y desambiguados, calcular el centroide geográfico (puntos 2 y 3 del algoritmo expuesto en el punto anterior).

Este centroide geográfico es utilizado para representar a cada documento, así entonces, la similitud o cercanía entre documentos es determinada por medio de medir distancias geográficas. Dado que se están considerando $n$ *documentos ejemplo* en el proceso del re-ordenamiento, para este tipo de representación en particular esto significará que se tienen $n$ centroides geográficos. El procedimiento para generar la lista re-ordenada bajo este tipo de representación geográfica fue: *1)* medir la distancia geográfica entre cada documento recuperado $r_i$ contra cada *documento ejemplo* $e_i$; *2)* asignar a $r_i$ el valor de distancia mínimo que se haya obtenido en el paso anterior; finalmente, *3)* este valor es empleado para generar el nuevo ordenamiento de la lista de documentos recuperados.

**Combinada (Temático y Geográfico).** Nos referiremos como tipo de representación *combinada* al hecho de mezclar tanto a la información proporcionada por los elementos temáticos de los documentos como a la información geográfica.

Para hacer esta combinación de fuentes de información empleamos la fórmula:

$$Ranking(q,d) = \lambda(Sim_{tem}(q,d)) + (1-\lambda)(Sim_{geo}(q,d)); \qquad (4.3.1)$$

donde la similitud temática ($Sim_{tem}$) es obtenida por medio de representar a los documentos con todas sus palabras(i.e., $BOW$) excepto las entidades geográficas. Así entonces, la similitud temática puede ser determinada por medio de medir similitudes entre vectores por medio de la fórmula del coseno. Por otro lado, la similitud geográfica ($Sim_{geo}$) es obtenida por medio de representar a los documentos ya sea a través de sólo entidades geográficas verificadas (i.e., representación *geográfica simple*), elementos geográficos expandidos (i.e., representación *geográfica expandida*) o por medio de la representación con *coordenadas geográficas*. Dependiendo de la forma de representación geográfica dependerá la forma en cómo se calculará la similitud geográfica.

Al final, ambos valores de similitud son combinados por medio de un factor $\lambda$ el cual determina la importancia que se le dará ya sea a la parte temática y/o a la parte geográfica.

## 4.4. Resumen

En resumen, en este capítulo hemos explicado la arquitectura propuesta para nuestra estrategia de re-ordenamiento. La idea principal de nuestra propuesta es emplear la información obtenida a través de una técnica de retroalimentación de relevancia para el proceso de re-ordenamiento y no para un proceso de reformulación de consultas, siendo este último el esquema típico en un sistema de IR.

En particular se propone el uso de documentos completos, denominados *documentos ejemplo* para realizar el re-ordenamiento de una lista de documentos. La razón principal para este hecho es debido a que suponemos que a través de *documentos ejemplo* es más fácil para el usuario expresar sus necesidades de información, las cuales muchas veces llevan consigo requerimientos de información en forma implícita, hecho que es frecuente en consultas geográficas.

El primer problema al que nos enfrentamos fue el cómo manejar la información de los documentos ejemplo. Para esto se propusieron tres formas distintas de manejar la información de los ejemplos: *i)* a través de un documento virtual, el cual en su

interior lleva toda la información de los ejemplos; *ii)* un resumen multi-documento, el cual contiene temas en común y diferencias importantes de los *documentos ejemplo*; y *iii)* a través de un MRF, donde se espera que tras iterar el campo combinando varias fuentes de información, incluyendo la intención de búsqueda (i.e., los *documentos ejemplo*), sea posible separar los elementos relevantes de los que no lo son.

Finalmente está el problema de la información que se utilizó para representar a los documentos. Para esto, siendo el caso simple, se propuso el emplear toda la información disponible, i.e., todas las *palabras* (BOW). Adicionalmente, tres formas de representación puramente geográfica fueron propuestas: *i)* entidades geográficas simples, *ii)* entidades geográficas expandidas, y *iii)* a través de coordenadas geográficas. Por último, una forma de representación combinada, i.e., donde elementos temáticos y geográficos son mezclados con el objetivo de evaluar qué tan complementarias son ambas formas de representación. En el capítulo siguiente se describen los experimentos realizados así como los resultados obtenidos con estas propuestas de solución.

# Capítulo 5

# Experimentos y Resultados

En este capítulo se describen los experimentos planteados para resolver el problema del re-ordenamiento por relevancia en el contexto de búsquedas geográficas. Como se ha mencionado en secciones anteriores, la máquina de IR no representa el foco principal de este trabajo, sin embargo, dentro de este capítulo se describe brevemente el sistema empleado como máquina de IR así como la configuración que se le dió. Agregado a esto, se describe también el conjunto de datos y las consultas empleadas para la realización de los experimentos.

En la primera sección, junto con la descripción de los datos se detalla en qué consiste la tarea del GeoCLEF, principal foro de evaluación para sistemas de recuperación de información geográfica. Posteriormente se describen en detalle los experimentos planteados de acuerdo a las diferentes formas definidas para el manejo de los *documentos ejemplo* así como de las distintas formas de representación de dichos documentos (Véase capítulo 4).

## 5.1. Máquina de IR Lemur

El *módulo de IR* contemplado en la arquitectura propuesta (figura 4.1) está formado por el sistema de recuperación de información LEMUR. El sistema LEMUR IR forma parte del Lemur Project[1] realizado por la Universidad de Massachusetts en colaboración con la Universidad de Carnegie Mellon. La colección de datos fue indexada utilizando este sistema; para todos los experimentos que se realizaron se creó un sólo índice. Para la creación de este índice, la colección de documentos fue pre-procesada, es decir, se hizo eliminación de palabras vacías así como también se

[1] http://www.lemurproject.org

aplicó un proceso de truncado de las palabras.

Las palabras vacías son aquellas que no aportan información, e.g., artículos, preposiciones o conjunciones. En particular se utilizó una lista de 571 palabras vacías proporcionada por el foro CLEF [1]. Mientras que por otro lado, el truncado de palabras tiene por objetivo llevar o transformar las palabras lo más cerca posible a su raíz léxica, en particular utilizamos el truncador de palabras Porter [76].

Como resultados base se tomaron los producidos por LEMUR bajo la configuración del modelo de espacio vectorial (VSM, por sus siglas en Inglés). Es conveniente mencionar que dado que LEMUR es un sistema que tiene varios años en constante desarrollo, actualmente está considerado como uno de los sistemas de IR más robustos [92], capaz de producir resultados adecuados para propósitos de comparación contra nuevos esquemas de recuperación.

## 5.2. GeoCLEF

El objetivo principal del GeoCLEF es proporcionar un marco de referencia con el cual evaluar sistemas de recuperación de información geográfica, los cuales desempeñan una tarea de búsqueda, la cuál involucra tanto aspectos espaciales como multilingües. Dentro del GeoCLEF, la tarea de recuperación de información geográfica es propuesta como una tarea de recuperación *ad hoc*, lo que significa que la recuperación se realiza sobre colecciones cerradas de documentos.

### 5.2.1. Conjunto de Datos

Durante los cuatro años en los que se realizó este ejercicio (2005-2008) el conjunto de datos utilizado está formado por noticias de los años 1994 y 1995, todas en idioma Inglés[2]. Las noticias contienen historias que cubren tanto eventos nacionales como internacionales, en consecuencia, contienen una gran variedad de referencias geográficas. La siguiente tabla (Tabla 5.2.1) muestra algunas estadísticas sobre la colección de datos.

[2]Además del Inglés, Alemán y Portugués son los idiomas para los que la tarea GIR se encuentra definida, para ver más detalles sobre las colecciones de datos utilizadas para estos dos últimos refiérase a [32, 31, 65].

| Nombre | Origen | Num. documentos | Idioma | Dominio |
|---|---|---|---|---|
| GH95 | *The Glasgow Herald* | 56,472 | Inglés | Noticias |
| LAT94 | *The Los Angeles Times* | 113,005 | Inglés | Noticias |
| | | Total: 169,477 | | |

**Tabla 5.1: Conjuntos de datos utilizados dentro del GeoCLEF**

### 5.2.2. Tópicos

Para cada ejercicio del GeoCLEF, se entregan 25 *tópicos* o consultas a los grupos participantes, hasta la fecha existen 100 tópicos disponibles para la tarea GIR. El título, o consulta principal, corresponde a lo que se encuentra entre las etiquetas <EN-title> y </EN-title>. Junto con ésta, se dan una breve descripción (<EN-desc>,</EN-desc>) y una narración (<EN-narr>,</EN-narr>), las cuales contienen información sobre los detalles de la búsqueda geográfica y el criterio de relevancia respectivamente.A continuación se muestra un ejemplo del formato dado para cada tópico.

```
<top>
<num>GC030</num>
<EN-title>Car bombings near Madrid</EN-title>
<EN-des>Documents about car bombings occurring near
Madrid</EN-desc>
<EN-narr>Relevant documents treat cases of car bombings
occurring in the capital of Spain and its outskirts</EN-narr>
</top>
```

Para los experimentos realizados se emplearon sólo aquellos tópicos que tuvieran más de 10 documentos relevantes en la colección de documentos, quedando así un total de 66 tópicos para la realización de nuestros experimentos. Este criterio se aplicó debido a que nuestra propuesta de solución parte del hecho de que las consultas tienen efectivamente elementos relevantes recuperados.

### 5.2.3. Evaluación

Dado que nuestros experimentos fueron realizados dentro de una colección cerrada de documentos nos es posible evaluar el desempeño de nuestro sistema por medio de las medidas $AveP$ y $MAP$ expuestas en el capítulo 2. Adicionalmente a estas medidas, también consideramos la medida *R-prec* y la medida $P@k$ para mostrar los resultados. Es conveniente mencionar que la medida MAP se consideró a mil documentos para todos los experimentos realizados. Cada resultado MAP va acompañado de un número entre paréntesis (+/-*dif* %) el cual indica el porcentaje de ganancia y/o perdida del resultado obtenido con respecto al método base.

Finalmente, se aplico una prueba de validación estadística (*paired students t-test*) para validar la significancia de los resultados obtenidos en los diferentes experimentos realizados. Para la realización de esta prueba nos fijamos en los cambios de la medida MAP. En las tablas de resultados se marcó con el símbolo asterisco (*****) a los resultados que logran invalidar la hipótesis nula con una confianza de al menos del 90 %.

## 5.3. Configuración de los Experimentos

Tres grandes bloques de experimentos fueron planteados para cumplir con los objetivos de este trabajo de tesis. La principal variante entre estos experimentos es la forma en cómo se manejó la información de los *documentos ejemplo*. Así pues, el primer conjunto de experimentos realizados maneja a los ejemplos a través de un "documento virtual" (Véase sección 5.4). El segundo conjunto de experimentos hace lo propio por medio de un "resumen multi-documento" (Véase sección 5.5). Finalmente, el tercer bloque de experimentos maneja la información de los *documentos ejemplo* junto con información externa proveniente de la máquina de IR por medio de un campo aleatorio de Markov (Véase sección 5.6).

Para cada uno de estos tres experimentos se consideraron dos formas distintas para la obtención de los *documentos ejemplo*, las cuales fueron: *i)* por medio de un proceso de selección ciega; y *ii)* por medio de un proceso de selección simulada. Véase la sección 2.1.3 para recordar en consisten estos dos tipos de selección de documentos.

Finalmente, para cada modalidad de selección se consideraron siete formas de representación: *BOW*, *geográfica simple*, *geográfica expandida*, *coordenadas geográficas*, *combinada (temática, geográfica simple)*, *combinada (temática, geográfica expandida)*

y *combinada (temática, coordenadas geográficas)*. Véase sección 4.3 para conocer en que consiste cada tipo de representación planteada.

## 5.4. Experimentos manejando un *Documento Virtual*

La idea de construir un "documento virtual" es con la finalidad de conservar la mayor información posible de los *documentos ejemplo*, por medio de este "documento virtual" se espera obtener una aproximación explícita de las necesidades de información del usuario, y posteriormente generar un orden más pertinente de la lista de documentos recuperados (Véase sección 4.2.1).

Los experimentos consistieron en seleccionar un número $n$ de *documentos ejemplo* por medio de una estrategia de retroalimentación y llevarlos a su forma de un documento virtual, el cual como se explicó en el capítulo anterior es el resultado de la unión de los $n$ *documentos ejemplo*. Una vez construido el documento virtual se re-ordenan los documentos recuperados de acuerdo al grado de cercanía que éstos tengan con el documento virtual.

Para medir esta cercanía se emplearon distintas medidas, las cuales dependen del tipo de representación que se le da a los documentos (Véase sección 4.3) ya sea una representación *BOW*, *geográfica simple*, *geográfica expandida*, *coordenadas geográficas* y/o *combinada (temática, geográfica)*. Finalmente, es conveniente mencionar que para todos los experimentos, $n$ se fijó a 2, 5 y 10 documentos, los cuales como veremos en secciones posteriores fueron seleccionados en dos modalidades: *ciega* y *simulada*.

El objetivo de los experimentos presentados en esta sección es mostrar que el uso de los *documentos ejemplo*, los cuales son obtenidos por medio del módulo de retroalimentación (Véase figura 4.1) permiten generar un orden más pertinente del conjunto de documentos recuperados por una máquina de IR en el contexto de búsquedas geográficas.

### 5.4.1. Retroalimentación Ciega

Dentro de este primer bloque de experimentos se muestran los resultados obtenidos con la estrategia propuesta para el re-ordenamiento manejando los *documentos ejemplo* a través de un documento virtual. En esta sección se considera un proce-

so de retroalimentación ciega, i.e., se considera a los primeros $n$ documentos como documentos relevantes y por lo tanto son tomados como *documentos ejemplo*.

La tabla 5.2 muestra los resultados obtenidos cuando se representó de forma $BOW$ tanto al documento virtual como los documentos recuperados. La última fila de la tabla muestra los resultados generados por Lemur, i.e., el método base. Los números resaltados en negritas indican la configuración en la que se ha superado al método base, mientras que los números entre paréntesis representan el porcentaje de la ganancia (+) o de la pérdida(-) del método propuesto con respecto al MAP generado por el método base.

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | 0.15 | 0.24 | 0.24(-3.1) |
| 5 *docs* | 0.15 | 0.24 | 0.24(-2.1) |
| 10 *docs* | 0.16 | 0.27 | **0.26**(+3.6) |
| *método base* | *0.16* | *0.27* | *0.25* |

**Tabla 5.2: Experimento manejando: documento virtual; con representación: *BOW*; selección ciega**

Como es posible observar, empleando una forma de representación $BOW$ sólo es posible superar el ordenamiento del método base cuando se emplean hasta diez *documentos ejemplo*, sin embargo, bajo esta misma configuración es posible ver que la precisión a cien ($P$@100) así como la medida *R-prec* se ven afectadas, siendo esto un indicador de que aunque se está mejorando el orden en el total de mil documentos, en las primeras posiciones de la lista de documentos recuperados se están removiendo algunos elementos relevantes. Este hecho provoca que aunque el MAP es mejor (+3.6 %) comparado con el obtenido por el método base, en realidad estos resultados no son significativos.

Los siguientes experimentos que se hicieron bajo este mismo esquema consideran únicamente la información geográfica de los *documentos ejemplo*. La tabla 5.3 muestra los resultados cuando el documento virtual y los documentos recuperados fueron representados sólo por medio de sus entidades geográficas, i.e., una forma de representación *geográfica simple*, mientras que la tabla 5.4 muestra los resultados obtenidos al considerar una forma de representación *geográfica expandida* y la tabla 5.5 muestra los resultados obtenidos al representar la información de los documentos por medio de *coordenadas geográficas.*

Es importante hacer notar que en los resultados obtenidos de estos experimentos en ningún caso es posible mejorar el orden generado por el método base. De estos

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | 0.12 | 0.19 | 0.18(-27) |
| 5 *docs* | 0.12 | 0.21 | 0.20(-20) |
| 10 *docs* | 0.13 | 0.24 | 0.22(-13) |
| *método base* | *0.16* | *0.27* | *0.25* |

**Tabla 5.3: Experimento manejando: documento virtual; con representación: *Geográfica Simple*; selección ciega**

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | 0.11 | 0.19 | 0.17(-30) |
| 5 *docs* | 0.12 | 0.21 | 0.19(-22) |
| 10 *docs* | 0.13 | 0.23 | 0.21(-15) |
| *método base* | *0.16* | *0.27* | *0.25* |

**Tabla 5.4: Experimento manejando: documento virtual; con representación: *Geográfica Expandida*; selección ciega**

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | 0.06 | 0.09 | 0.08(-67) |
| 5 *docs* | 0.07 | 0.12 | 0.11(-55) |
| 10 *docs* | 0.07 | 0.17 | 0.14(-42) |
| *método base* | *0.16* | *0.27* | *0.25* |

**Tabla 5.5: Experimento manejando: documento virtual; con representación: *Coordenadas Geográficas*; selección ciega**

experimentos el que tiene un mejor comportamiento es para el caso de representación *geográfica simple* (Tabla 5.3),pues es en el cual las perdidas con respecto al método base son menores. Cuando se hace una representación *geográfica expandida* (Tabla 5.4) el desempeño del sistema empieza a decaer aún más, lo cual nos da indicios para pensar que la expansión está agregando ruido a los documentos, i.e., términos geográficos que no aportan información discriminante para el proceso de re-ordenamiento y además que algunos de los documentos seleccionados como ejemplos están desviando el foco geográfico de la consulta original.

Por otro lado, para el caso de un tipo de representación por medio de *coordenadas geográficas* (Tabla 5.5) se obtienen los resultados más bajos, lo cual nos indica que los focos geográficos de los *documentos ejemplo* no se encuentran concentrados y en su lugar se encuentran muy dispersos entre si.

Los resultados tan bajos que se obtuvieron con estas tres formas de representación geográfica indican que la información geográfica de los documentos por si sola no es suficiente para poder producir un ordenamiento adecuado, y más bien es información complementaria. Es principalmente por esta razón que se propone una forma de

representación *combinada (temática y geográfica)* para hacer el ordenamiento de los documentos.

Las tablas siguientes muestran los resultados al hacer esta combinación de fuentes de información por medio de la fórmula 4.3.1 descrita en el método propuesto (Véase sección 4.3). La tabla 5.6 muestra los resultados de esta combinación cuando la información geográfica considerada para la forma de representación *combinada* es únicamente las entidades geográficas originalmente contenidas en los documentos; en otras palabras el equivalente a la representación *geográfica simple*. Por otro lado, la tabla 5.7 muestra los resultados obtenidos al representar la fuente de información geográfica por medio de sus elementos *geográficos expandidos*. Finalmente, la tabla 5.8 muestra los resultados de esta combinación, cuando la información geográfica es representada por medio de *coordenadas geográficas*.

Recordemos que la fórmula 4.3.1 involucra el uso de un factor *lambda*, el cual define a cuál de los términos de la fórmula se le dará más importancia. Mientras mayor sea el factor *lambda* quiere decir que mayor importancia se le está dando a la información temática de los documentos.

| *num docs* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| 2 *docs* | 0.13 | 0.21 | 0.21(-15) | 0.15 | 0.23 | 0.24(-4.9) | 0.15 | 0.25 | **0.25**(+0.8) | 0.16 | 0.25 | 0.25(-0.9) |
| 5 *docs* | 0.13 | 0.23 | 0.22(-11) | 0.15 | 0.26 | 0.24(-2.2) | 0.16 | 0.27 | **0.25**(+1.8) | 0.16 | 0.26 | 0.25(-0.2) |
| 10 *docs* | 0.14 | 0.26 | 0.24(-5.7) | 0.16 | 0.27 | **0.25**(+0.7) | **0.16** | **0.28** | **0.26**(+4.7)* | 0.16 | **0.27** | **0.26**(+4.4)* |
| *método base: p@100 = 0.16 ; R-prec = 0.27 ; MAP = 0.25* | | | | | | | | | | | | |

**Tabla 5.6: Experimento manejando: documento virtual; con representación: *Temática y Geográfica Simple*; selección ciega**

| *num docs* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| 2 *docs* | 0.13 | 0.20 | 0.20(-20) | 0.14 | 0.23 | 0.22(-10) | 0.15 | 0.25 | 0.25(-1.8) | 0.16 | 0.24 | 0.24(-2.7) |
| 5 *docs* | 0.13 | 0.23 | 0.22(-11) | 0.15 | 0.26 | 0.24(-3.9) | 0.16 | 0.26 | **0.25**(+1) | 0.16 | 0.26 | 0.25(-0.9) |
| 10 *docs* | 0.14 | 0.25 | 0.23(-7.6) | 0.16 | 0.27 | **0.25**(+0.2) | **0.17** | **0.27** | **0.26**(+4.6)* | 0.16 | 0.27 | **0.26**(+3.9) |
| *método base: p@100 = 0.16 ; R-prec = 0.27 ; MAP = 0.25* | | | | | | | | | | | | |

**Tabla 5.7: Experimento manejando: documento virtual; con representación: *Temática y Geográfica Expandida*; selección ciega**

Para estos experimentos, es decir, combinando la información temática y geográfica, los mejores resultados se obtienen en la tabla 5.6, los cuales corresponden a la combinación de la información temática representada a través de vectores de palabras y de la información geográfica representada a través de vectores de entidades

| *num docs* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| 2 *docs* | 0.15 | 0.23 | 0.23(-8.7) | 0.15 | 0.23 | 0.23(-8.7) | 0.15 | 0.23 | 0.23(-8.7) | 0.15 | 0.23 | 0.23(-8.3) |
| 5 *docs* | 0.15 | 0.24 | 0.24(-5.4) | 0.15 | 0.24 | 0.24(-5.4) | 0.15 | 0.24 | 0.24(-5.4) | 0.15 | 0.24 | 0.24(-5.2) |
| 10 *docs* | 0.15 | 0.26 | 0.25(+0.8) | 0.15 | 0.26 | **0.25**(+0.8) | 0.15 | 0.26 | **0.25**(+0.8) | 0.15 | 0.27 | **0.25**(+1) |
| *método base: p@100 = 0.16 ; R-prec = 0.27 ; MAP = 0.25* | | | | | | | | | | | | |

**Tabla 5.8: Experimento manejando: documento virtual; con representación: *Temática y Coordenadas Geográficas*; selección ciega**

geográficas, i.e., una representación *geográfica simple*. Nótese que es en esta misma tabla, con $\lambda$ igual a 0.6 y considerando 10 *documentos ejemplo*, cuando el mejor desempeño es obtenido, siendo capaz de superar el resultado obtenido por el método base en las tres medidas expuestas. Además de esto, gracias a la prueba de validación estadística nos es posible afirmar que estos resultados son en efecto importantes y no causa del azar.

Como nota final de este bloque de experimentos es importante recalcar que como factor común en la forma de representación combinada (i.e., Tablas 5.6, 5.7, y 5.8), sucede que cuando $\lambda$ tiende a ser mayor ($\lambda > 0.4$), mejores resultados se obtienen en las diferentes configuraciones, lo cual sugiere que la información temática influye fuertemente en la discriminación entre documentos relevantes y no relevantes.

### 5.4.2. Retroalimentación Simulada

Los experimentos realizados en esta sección, similares a los presentados anteriormente, consideran el manejo de la información de los *documentos ejemplo* por medio de la construcción de un documento virtual. La única diferencia es que para estos experimentos se consideró una estrategia de retroalimentación simulada, es decir, la relevancia de los *documentos ejemplo* es determinada por un usuario en forma manual.

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | *0.16* | *0.29* | *0.29* |
| 5 *docs* | *0.17* | *0.36* | *0.37* |
| 10 *docs* | *0.18* | *0.48* | *0.50* |

**Tabla 5.9: Desempeño del sistema tras colocar al principio de la lista a los documentos seleccionados de forma simulada**

Para estos experimentos consideramos como punto de comparación los resultados que se obtienen tras aplicar el proceso de retroalimentación de relevancia en forma simulada, es decir, los valores que se obtienen tras colocar al principio de la lista los

documentos otorgados por el proceso de retroalimentación sin aplicar la estrategia de re-ordenamiento. La tabla 5.9 muestra el desempeño del sistema bajo este esquema. En las tablas de resultados sólo se remarcan en negritas sólo aquellos resultados que mejoran a los obtenidos tras sólo aplicar la retroalimentación simulada, es decir a los mostrados en la tabla 5.9. Es conveniente mencionar que para los experimentos realizados con un proceso de selección simulada, las tablas de resultados muestran la ganancia y/o pérdida del método propuesto (+/-*dif* %) en comparación con los MAPs obtenidos en la tabla 5.9.

La tabla 5.10 muestra los resultados que se obtienen al manejar la representación *BOW* para el documento virtual así como para los documentos recuperados. Los resultados muestran que con tan sólo dos documentos entregados por el proceso de retroalimentación de relevancia en forma simulada es posible lograr un desempeño que supera tanto al método base como al proceso único de la retroalimentación (Tabla 5.9).

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | **0.18** | **0.36** | **0.36**(+25)* |
| 5 *docs* | **0.20** | **0.42** | **0.45**(+21)* |
| 10 *docs* | **0.21** | **0.53** | **0.57**(+13)* |

**Tabla 5.10: Experimento manejando: documento virtual; con representación: *BOW*; selección simulada**

Las tablas 5.11, 5.12 y 5.13 muestran los resultados obtenidos cuando el documento virtual y los documentos recuperados son representados en su forma *geográfica simple*, *geográfica expandida* y con *coordenadas geográficas* respectivamente.

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | 0.14 | 0.28 | 0.27(-4.9) |
| 5 *docs* | 0.16 | **0.37** | **0.39**(+5.3) |
| 10 *docs* | **0.18** | **0.50** | **0.53**(+5)* |

**Tabla 5.11: Experimento manejando: documento virtual; con representación: *Geográfica Simple*; selección simulada**

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | 0.13 | 0.27 | 0.27(-7.2) |
| 5 *docs* | 0.15 | **0.37** | **0.39**(+4) |
| 10 *docs* | 0.18 | **0.49** | **0.52**(+4.2)* |

**Tabla 5.12: Experimento manejando: documento virtual; con representación: *Geográfica Expandida*; selección simulada**

Para el caso de las representaciones geográficas, los mejores resultados se dan en

la tabla 5.11, que son los experimentos que corresponden a representar los documentos por medio de la forma *geográfica simple*, aunque como es posible observar, para superar a los resultados obtenidos tras sólo aplicar el proceso de retroalimentación en las tres medidas expuestas, es necesario considerar hasta diez *documentos ejemplo*.

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | 0.07 | 0.14 | 0.14(-50) |
| 5 *docs* | 0.10 | 0.26 | 0.27(-27) |
| 10 *docs* | 0.13 | 0.42 | 0.42(-16) |

**Tabla 5.13: Experimento manejando: documento virtual; con representación: *Coordenadas Geográficas*; selección simulada**

De manera similar a los resultados de la sección anterior, es posible observar que la representación *geográfica expandida* (Tabla 5.12) y la de *coordenadas geográficas* (Tabla 5.13) resultan en un degradamiento del desempeño del sistema, siendo éste último el caso donde no es posible superar en ninguna de las medidas expuestas al proceso de sólo aplicar la retroalimentación de relevancia simulada. Estos resultados confirman nuestra suposición inicial, en la cual se decía que el manejar sólo elementos geográficos no permite hacer un adecuado ordenamiento pues son elementos muy dispersos entre si. En estos experimentos esto es más claro, pues aunque son *documentos ejemplo* relevantes, pues fueron seleccionados de manera simulada, sus elementos geográficos son muy variados de tal forma que no es posible obtener ganancias significativas el ordenamiento e incluso en ocasiones se pierde mucho, tal es el caso de la tabla 5.13.

Las siguientes tablas (5.14, 5.15, y 5.16) muestran los resultados al representar a los documentos por medio de una combinación de la información temática y geográfica. Por un lado, la tabla 5.14 muestra los resultados de esta combinación cuando la información geográfica considerada son únicamente las entidades geográficas originalmente contenidas en los documentos, i.e., la forma *geográfica simple*. Por otra parte, la tabla 5.15 muestra los resultados obtenidos al representar la fuente de información geográfica por medio de su forma *geográfica expandida*. Y, finalmente la tabla 5.16 muestra los resultados de esta combinación, cuando la información geográfica es representada por medio de *coordenadas geográficas*.

En estos experimentos, casi bajo todas las configuraciones es posible superar el desempeño del proceso único de retroalimentación (Tabla 5.9). También es importante hacer notar que el mismo fenómeno que cuando se consideró un proceso de retroalimentación ciega se sigue dando, es decir, que a valores mayores de $\lambda$ se obtie-

| *num* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *docs* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| 2 *docs* | 0.15 | 0.30 | **0.31**(+8.2)* | 0.17 | **0.34** | **0.35**(+20)* | **0.18** | **0.36** | **0.37**(+27)* | **0.18** | **0.36** | **0.36**(+26)* |
| 5 *docs* | 0.17 | **0.40** | **0.42**(+13)* | **0.18** | **0.42** | **0.45**(+20)* | **0.20** | **0.44** | **0.46**(+24)* | **0.20** | **0.44** | **0.46**(+23)* |
| 10 *docs* | **0.19** | **0.51** | **0.55**(+8.8)* | **0.20** | **0.53** | **0.56**(+12)* | **0.21** | **0.54** | **0.57**(+15)* | **0.21** | **0.54** | **0.57**(+14)* |

**Tabla 5.14: Experimento manejando: documento virtual; con representación: *Temática y Geográfica Simple*; selección simulada**

| *num* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *docs* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| 2 *docs* | 0.15 | **0.30** | **0.31**(+6.5) | **0.17** | **0.35** | **0.35**(+21)* | **0.18** | **0.36** | **0.37**(+28)* | **0.18** | **0.35** | **0.36**(+25)* |
| 5 *docs* | **0.17** | **0.40** | **0.42**(+13)* | **0.19** | **0.44** | **0.45**(+21)* | **0.20** | **0.44** | **0.46**(+24)* | **0.20** | **0.43** | **0.46**(+22)* |
| 10 *docs* | **0.19** | **0.51** | **0.54**(+8.2)* | **0.20** | **0.53** | **0.56**(+13)* | **0.21** | **0.54** | **0.57**(+14)* | **0.21** | **0.54** | **0.57**(+14)* |

**Tabla 5.15: Experimento manejando: documento virtual; con representación: *Temática y Geográfica Expandida*; selección simulada**

nen mejores resultados, siendo en $\lambda$ =0.6 el punto donde parecen alcanzar el mejor compromiso las configuraciones que consideran las entidades *geográficas simples* (Tabla 5.14) y las *geográficas expandidas* (Tabla 5.15).

| *num* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *docs* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| 2 *docs* | **0.18** | **0.33** | **0.34**(+17)* | **0.18** | **0.33** | **0.34**(+17)* | **0.18** | **0.33** | **0.34**(+17)* | **0.18** | **0.33** | **0.34**(+17)* |
| 5 *docs* | **0.19** | **0.41** | **0.43**(+16)* | **0.19** | **0.41** | **0.43**(+16)* | **0.19** | **0.41** | **0.43**(+16)* | **0.19** | **0.41** | **0.43**(+16)* |
| 10 *docs* | **0.21** | **0.52** | **0.55**(+10)* | **0.20** | **0.52** | **0.55**(+10)* | **0.20** | **0.52** | **0.55**(+10)* | **0.21** | **0.52** | **0.56**(+11)* |

**Tabla 5.16: Experimento manejando: documento virtual; con representación: *Temática y Coordenadas Geográficas*; selección simulada**

Es importante recalcar que para estos experimentos, la combinación de la información temática con los elementos *geográficos expandidos* (Tabla 5.15) permite obtener un mejor desempeño similar al obtenido con la combinación de la información temática con las entidades *geográficas simples* (Tabla 5.14), indicador de que el proceso de expansión geográfica se está haciendo de manera correcta. Nótese también que la combinación de la información temática con *coordenadas geográficas* (Tabla 5.16) sigue sin aportar información que ayude a discriminar entre los elementos relevantes y los no relevantes.

Finalmente es importante mencionar que cuando se hace la combinación de los elementos temáticos y geográficos a través de la fórmula 4.3.1 es cuando en general se obtienen mejores resultados, superando tanto a la forma de representación *BOW* así como a las tres variantes de representación geográfica, mostrando con esto la complementariedad de estos elementos.

## 5.5. Experimentos manejando un *Resumen Multi-Documento*

Como se mencionó en el capítulo anterior, la idea de manejar resúmenes surge principalmente por la necesidad de eliminar información contenida en los *documentos ejemplo* que no está aportando elementos suficientes para hacer un ordenamiento adecuado. El objetivo de hacer resúmenes es el de concentrar la información similar de los *documentos ejemplo* y dejar de lado información contradictoria o poco relevante, con esto se espera también que el número de entidades geográficas contenidas en los documentos se reduzca a sólo las entidades geográficas realmente relevantes.

Para estos experimentos se manejó la información de los *documentos ejemplo* por medio de construir un resumen multi-documento (Véase sección 4.2.2). Al igual que para los experimentos anteriores los *n documentos ejemplo* son seleccionados tanto de manera ciega como simulada y también se consideraron 2, 5 y 10 documentos en el proceso de retroalimentación.

En estos experimentos, tras obtener los ejemplos, se aplica el proceso de generación de resumen multi-documento explicado en la sección 4.2.2. Una vez construido este resumen multi-documento, para el cual en todos los experimentos se fijó su razón de compresión al 40 % , éste es comparado contra cada uno de los elementos de la lista de documentos recuperados de tal forma que permita hacer un re-ordenamiento de éstos.

De manera similar al conjunto de experimentos realizados en la sección anterior, cuando se manejó la información de los *documentos ejemplo* a través de un documento virtual, aquí también dividimos los conjuntos de experimentos considerando primero un modo de retroalimentación ciega y posteriormente un modo de retroalimentación simulada.

### 5.5.1. Retroalimentación Ciega

La tabla 5.17 muestra los resultados obtenidos cuando el resumen multi-documento es representado a través de su forma *BOW* o de bolsa de palabras.

Para estos experimentos sólo es posible mejorar el MAP cuando hasta 10 *documentos ejemplo* son considerados para el resumen multi-documento. Note que este MAP es menor con respecto a su experimento equivalente mostrado en la tabla 5.2,

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | 0.16 | 0.24 | 0.24(-5.2) |
| 5 *docs* | 0.16 | 0.25 | 0.24(-2.5) |
| 10 *docs* | 0.16 | 0.27 | **0.26**(+3.2) |
| *método base* | *0.16* | *0.27* | *0.25* |

**Tabla 5.17: Experimento manejando: resumen multi-documento; con representación: *BOW*; selección ciega**

sin embargo, los valores obtenidos para la medida *R-prec* manejando un resumen multi-documento y a partir de que se consideran 5 *documentos ejemplo*, son mejores que los que se obtienen cuando se maneja el documento virtual (Tabla 5.2). En otras palabras, el resumen multi-documento no mejora el orden de los mil documentos, pero sí permite colocar más documentos relevantes en las primeras $R$ posiciones. Esto muestra hasta cierto punto que nuestra hipótesis inicial sobre el manejo de resúmenes se esta logrando, pues el resumen logra concentrar de mejor manera la información discriminativa, sin embargo siguen sin ser resultados significativos.

Las siguientes tablas (5.18, 5.19, y 5.20) muestran los resultados obtenidos al representar el resumen multi-documento por medio de sólo elementos geográficos. La tabla 5.18 considera la representación *geográfica simple*, la tabla 5.19 una forma de representación *geográfica expandida* y la tabla 5.20 una representación por medio de *coordenadas geográficas*.

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | 0.10 | 0.18 | 0.17(-33) |
| 5 *docs* | 0.11 | 0.20 | 0.19(-26) |
| 10 *docs* | 0.12 | 0.23 | 0.21(-17) |
| *método base* | *0.16* | *0.27* | *0.25* |

**Tabla 5.18: Experimento manejando: resumen multi-documento; con representación: *Geográfica Simple*; selección ciega**

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | 0.10 | 0.17 | 0.16(-36) |
| 5 *docs* | 0.11 | 0.20 | 0.18(-27) |
| 10 *docs* | 0.12 | 0.23 | 0.20(-19) |
| *método base* | *0.16* | *0.27* | *0.25* |

**Tabla 5.19: Experimento manejando: resumen multi-documento; con representación: *Geográfica Expandida*; selección ciega**

Note que en ningún caso fue posible superar el desempeño del método base con este tipo de representaciones. Sin embargo, de manera similar a los resultados obte-

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | 0.06 | 0.08 | 0.08(-68) |
| 5 *docs* | 0.07 | 0.12 | 0.11(-55) |
| 10 *docs* | 0.09 | 0.18 | 0.16(-37) |
| *método base* | *0.16* | *0.27* | *0.25* |

**Tabla 5.20: Experimento manejando: resumen multi-documento; con representación: *Coordenadas Geográficas*; selección ciega**

nidos empleando un documento virtual, la representación geográfica que da mejores resultados es la que considera sólo a las entidades originales del documento (Tabla 5.18) mientras que el agregar los términos expandidos (Tabla 5.19) y/o manejar coordenadas geográficas (Tabla 5.20) resulta en un degradamiento del desempeño del sistema. Esto sucede por la misma razón que el manejar un documento virtual con una representación puramente geográfica no obtiene buenos resultados, y es debido a la diversidad de los elementos geográficos contenidos en los *documentos ejemplo*. Y si a esto le agregamos que el resumen ha eliminado información, el desempeño de estos experimentos (Tablas 5.18, 5.19 y 5.20) comparado con el desempeño obtenido al manejar un documento virtual (Tablas 5.3, 5.4 y 5.5) es aún menor, lo cual se ve reflejado en los valores de las diferencias de cada experimento.

La tablas siguientes muestran los resultados obtenidos al hacer la combinación de fuentes de información por medio de la fórmula 4.3.1 descrita en el capítulo anterior (Véase sección 4.3). La tabla 5.21 muestra los resultados de esta combinación cuando la información geográfica considerada son únicamente las entidades geográficas originalmente contenidas en los documentos. Por otro lado, la tabla 5.22 muestra los resultados obtenidos al representar la fuente de información geográfica por medio de sus elementos geográficos expandidos. Finalmente, la tabla 5.23 muestra los resultados de esta combinación, cuando la información geográfica es representada por medio de coordenadas geográficas.

Los mejores resultados se obtienen en la tabla 5.21, los cuales corresponden a la combinación de la información temática representada a través de vectores de palabras y de la información geográfica representada a través de vectores de entidades geográficas simples. Nótese que es en esta misma tabla, con $\lambda$ igual a 0.8 y considerando 10 *documentos ejemplo*, cuando el mejor desempeño es obtenido, siendo capaz de superar el resultado del proceso único de retroalimentación en las tres medidas expuestas.

En estos experimentos (Tablas 5.21, 5.22, y 5.23), al igual que con sus equivalentes al manejar un documento virtual, sucede que cuando $\lambda$ tiende a ser mayor ($\lambda > 0.4$),

| *num* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *docs* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| 2 *docs* | 0.12 | 0.20 | 0.20(-21) | 0.14 | 0.23 | 0.22(-11) | 0.15 | 0.25 | 0.24(-3.9) | 0.16 | 0.25 | 0.24(-4.4) |
| 5 *docs* | 0.12 | 0.22 | 0.21(-16) | 0.14 | 0.25 | 0.23(-6.5) | 0.15 | 0.26 | 0.25(-1) | 0.16 | 0.26 | 0.25(-1.2) |
| 10 *docs* | 0.14 | 0.24 | 0.23(-9.5) | 0.15 | 0.26 | 0.25(-1.3) | **0.16** | 0.27 | **0.26**(+4)* | **0.17** | **0.28** | **0.26**(+4)* |
| *método base: p@100 = 0.16 ; R-prec = 0.27 ; MAP = 0.25* | | | | | | | | | | | | |

**Tabla 5.21: Experimento manejando: resumen multi-documento; con representación: *Temática y Geográfica Simple*; selección ciega**

| *num* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *docs* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| 2 *docs* | 0.12 | 0.18 | 0.18(-27) | 0.14 | 0.22 | 0.21(-15) | 0.15 | 0.24 | 0.23(-7) | 0.16 | 0.24 | 0.23(-6.7) |
| 5 *docs* | 0.12 | 0.22 | 0.21(-17) | 0.15 | 0.24 | 0.23(-6.2) | 0.15 | 0.25 | 0.25(-1.6) | 0.16 | 0.25 | 0.24(-2.2) |
| 10 *docs* | 0.14 | 0.25 | 0.22(-11) | 0.16 | 0.27 | 0.25(-1.7) | 0.16 | 0.26 | **0.25**(+0.2) | **0.16** | **0.27** | **0.26**(+3.3) |
| *método base: p@100 = 0.16 ; R-prec = 0.27 ; MAP = 0.25* | | | | | | | | | | | | |

**Tabla 5.22: Experimento manejando: resumen multi-documento; con representación: *Temática y Geográfica Expandida*; selección ciega**

| *num* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *docs* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| 2 *docs* | 0.15 | 0.23 | 0.22(-12) | 0.14 | 0.21 | 0.21(-16) | 0.15 | 0.23 | 0.22(-12) | 0.15 | 0.23 | 0.22(-12) |
| 5 *docs* | 0.15 | 0.24 | 0.23(-6.5) | 0.15 | 0.23 | 0.23(-8) | 0.15 | 0.24 | 0.23(-6.5) | 0.15 | 0.24 | 0.23(-6.4) |
| 10 *docs* | 0.15 | 0.26 | 0.25(-0.6) | 0.16 | 0.26 | 0.25(-0.7) | 0.15 | 0.26 | 0.25(-0.2) | 0.15 | 0.26 | 0.25(-0.1) |
| *método base: p@100 = 0.16 ; R-prec = 0.27 ; MAP = 0.25* | | | | | | | | | | | | |

**Tabla 5.23: Experimento manejando: resumen multi-documento; con representación: *Temática y Coordenadas Geográficas*; selección ciega**

mejores resultados se obtienen en las diferentes configuraciones, lo cual quiere decir que la información temática es de gran importancia para la discriminación entre documentos relevantes y no relevantes.

Es importante mencionar en este punto que para la mayoría de los casos cuando $\lambda = 0.8$ en las tablas 5.22, y 5.23 sucede que la medida *R-prec* es mejor comparada con los resultados obtenidos en sus experimentos equivalentes empleando un documento virtual (tablas 5.7 y 5.8), lo cual indica que efectivamente el resumen generado está ayudando a eliminar algunos términos geográficos irrelevantes contenidos dentro de los documentos así como también elementos temáticos poco discriminativos, permitiendo colocar a más documentos relevantes dentro de las primeras posiciones, pero no suficientes como para mejorar el orden global, lo cual se ve reflejado en la medida MAP y por tanto no significan una mejora significativa.

### 5.5.2. Retroalimentación Simulada

Los experimentos realizados en esta sección, similares a los presentados anteriormente, consideran el manejo de la información de los *documentos ejemplo* por medio de la construcción de un resumen multi-documento. La diferencia es que para estos experimentos se consideró una estrategia de retroalimentación simulada.

Tal y como se hizo en la sección 5.4.2, para estos experimentos tomamos como punto de comparación el desempeño del sistema tras aplicar el proceso de retroalimentación sin aplicar el proceso del re-ordenamiento. La tabla 5.9 muestra estos valores y de manera similar a los experimentos mostrados en aquella sección, aquí también se remarcan en negritas sólo los resultados que superan al proceso único de retroalimentación y las diferencias (positivas o negativas) mostradas en las tablas son con respecto a los resultados de la tabla 5.9.

La tabla 5.24 muestra los resultados obtenidos al manejar una forma de representación *BOW* para el resumen multi-documento así como para el conjunto de documentos recuperados. Como es posible observar, con tan sólo dos *documentos ejemplo* es posible obtener resultados que superan al proceso único de la retroalimentación de relevancia simulada (tabla 5.9).

Aunque los resultados obtenidos en la tabla 5.24 son buenos y significativos, es importante hacer notar que cuando se manejó un documento virtual bajo esta misma configuración (tabla 5.10) los resultados obtenidos son aún mejores. Este hecho

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | **0.18** | **0.33** | **0.34**(+18)* |
| 5 *docs* | **0.19** | **0.41** | **0.44**(+17)* |
| 10 *docs* | **0.20** | **0.52** | **0.56**(+11)* |

**Tabla 5.24: Experimento manejando: resumen multi-documento; con representación: *BOW*; selección simulada**

muestra dos cosas: *i)* que la estrategia de generación de resúmenes multi-documento permite en efecto generar documentos cortos (resúmenes) representativos del conjunto inicial de documentos, sin embargo, *ii)* se está sacrificando información que permitiría generar un orden igual de pertinente al que se obtiene cuando no se elimina ningún tipo de información, i.e., a cuando se maneja un documento virtual.

Las tablas 5.25, 5.26 y 5.27 muestran los resultados cuando el resumen multi-documento y los documentos recuperados son representados en su forma *geográfica simple, geográfica expandida* y por medio de *coordenadas geográficas* respectivamente.

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | 0.13 | 0.26 | 0.25(-13) |
| 5 *docs* | 0.15 | 0.34 | 0.36(-3.2) |
| 10 *docs* | 0.17 | **0.49** | **0.52**(+2.8)* |

**Tabla 5.25: Experimento manejando: resumen multi-documento; con representación: *Geográfica Simple*; selección simulada**

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | 0.12 | 0.24 | 0.25(-14) |
| 5 *docs* | 0.15 | 0.34 | 0.36(-3.9) |
| 10 *docs* | 0.17 | **0.49** | **0.51**(+1.8) |

**Tabla 5.26: Experimento manejando: resumen multi-documento; con representación: *Geográfica Expandida*; selección simulada**

| *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|
| 2 *docs* | 0.07 | 0.14 | 0.14(-51) |
| 5 *docs* | 0.09 | 0.25 | 0.26(-29) |
| 10 *docs* | 0.13 | 0.42 | 0.44(-12) |

**Tabla 5.27: Experimento manejando: resumen multi-documento; con representación: *Coordenadas Geográficas*; selección simulada**

De manera similar a los experimentos realizados empleando un documento virtual y una forma de retroalimentación simulada, aquí también sucede que los mejores resultados se dan cuando la forma de representación es por medio de las entidades

geográficas simples (tabla 5.25), aunque a diferencia de aquellos resultados (tabla 5.11), aquí sólo es posible superar al método base y al proceso de retroalimentación únicamente tras considerar 10 *documentos ejemplo* con la excepción de la medida *p@100*.

Además de esto, es posible también notar que el hacer una expansión de los elementos geográficos no resulta en un beneficio al momento de realizar el re-ordenamiento de los documentos, pues como se puede ver en la tabla 5.26 los resultados obtenidos son menores a los logrados con sólo las entidades geográficas simples (tabla 5.25).

Finalmente, es importante hacer notar que cuando manejamos las coordenadas geográficas como forma de representación (tabla 5.27), aunque no se logra superar al método base ni al proceso de retroalimentación, sí se logra obtener un mejor desempeño en comparación con su experimento equivalente manejando un documento virtual (tabla 5.13), en particular para el caso de cuando se otorgan 10 *documentos ejemplo*, lo cual es un indicativo de que efectivamente el resumen está ayudando a eliminar elementos geográficos irrelevantes contenidos en los documentos.

Las siguientes tablas (5.28, 5.29, y 5.30) muestran los resultados de representar tanto a los resúmenes multi-documento como al conjunto de documentos recuperados por medio de una combinación de la información temática y geográfica. Por un lado, la tabla 5.28 muestra los resultados de esta combinación cuando la información geográfica considerada es únicamente las entidades geográficas simples. Por otra parte, la tabla 5.29 muestra los resultados obtenidos al representar la fuente de información geográfica por medio de sus elementos geográficos expandidos. Y, finalmente la tabla 5.30 muestra los resultados de esta combinación, cuando la información geográfica es representada por medio de coordenadas geográficas.

| *num* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *docs* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| 2 *docs* | 0.14 | 0.28 | 0.29(+0.7) | 0.16 | **0.32** | **0.32**(+11)* | **0.17** | **0.33** | **0.34**(+19)* | **0.18** | **0.34** | **0.34**(+19)* |
| 5 *docs* | 0.17 | **0.37** | **0.39**(+5.6) | **0.18** | **0.39** | **0.42**(+13)* | **0.19** | **0.41** | **0.44**(+16)* | **0.20** | **0.42** | **0.44**(+18)* |
| 10 *docs* | **0.18** | **0.50** | **0.53**(+6.5)* | **0.20** | **0.52** | **0.55**(+9.7)* | **0.20** | **0.53** | **0.56**(+12)* | **0.21** | **0.53** | **0.56**(+12)* |

**Tabla 5.28: Experimento manejando: resumen multi-documento; con representación: *Temática y Geográfica Simple*; selección simulada**

Como es posible observar en estos experimentos, casi bajo todas las configuraciones es posible superar el desempeño del proceso único de retroalimentación de la tabla 5.9. También es importante hacer notar que el mismo fenómeno que se ha observado en experimentos anteriores se sigue dando, el cual es: que a valores mayores de $\lambda$, en

| *num docs* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| 2 *docs* | 0.14 | 0.27 | 0.28(-3.7) | 0.16 | **0.31** | **0.32**(+10)* | **0.18** | **0.34** | **0.34**(+18)* | **0.18** | **0.33** | **0.34**(+17)* |
| 5 *docs* | 0.16 | 0.36 | **0.39**(+4.6) | **0.18** | **0.40** | **0.42**(+13)* | **0.19** | **0.42** | **0.44**(+18)* | **0.19** | **0.41** | **0.44**(+17)* |
| 10 *docs* | **0.18** | **0.50** | **0.53**(+5.6)* | **0.20** | **0.52** | **0.55**(+10)* | **0.21** | **0.53** | **0.56**(+12)* | **0.20** | **0.53** | **0.56**(+11)* |

**Tabla 5.29: Experimento manejando: resumen multi-documento; con representación: *Temática y Geográfica Expandida*; selección simulada**

| *num docs* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| 2 *docs* | 0.17 | **0.31** | **0.32**(+10)* | 0.17 | **0.31** | **0.32**(+10)* | 0.17 | **0.31** | **0.32**(+10)* | **0.17** | **0.31** | **0.32**(+10)* |
| 5 *docs* | **0.18** | **0.40** | **0.42**(+12)* | **0.18** | **0.40** | **0.42**(+12)* | **0.18** | **0.40** | **0.42**(+12)* | **0.18** | **0.40** | **0.42**(+12)* |
| 10 *docs* | **0.20** | **0.51** | **0.54**(+8.3)* | **0.20** | **0.51** | **0.54**(+8.3)* | **0.20** | **0.51** | **0.54**(+8.3)* | **0.20** | **0.51** | **0.54**(+8.4)* |

**Tabla 5.30: Experimento manejando: resumen multi-documento; con representación: *Temática y Coordenadas Geográficas*; selección simulada**

particular $\lambda$ =0.6, mejores resultados se logran obtener para los casos de las tablas 5.28 y tabla 5.29, lo cual significa que la parte temática de los documentos sigue jugando un papel importante en el proceso del re-ordenamiento.

Al igual que en experimentos anteriores, también sucede que la combinación de la información temática con *coordenadas geográficas* (tabla 5.30) obtiene su mejor desempeño cuando $\lambda = 0.8$, indicando esto que las coordenadas geográficas aportan muy poca información. Esto sucede debido a que aunque se están eliminando términos geográficos, sigue habiendo una gran dispersión de los términos en el resumen multi-documento.

Finalmente es importante mencionar que aunque se obtienen resultados aceptables, estos siguen siendo ligeramente menores a los obtenidos cuando los *documentos ejemplo* son manejados a través de un documento virtual. Esto nos hace pensar que quizás si el tamaño de los resúmenes fuera un poco mayor, sería posible igualar los resultados de manejar documentos completos, pues recordemos que se esta manejando una razón de compresión del 40 %.

## 5.6. Experimentos manejando *Campos Aleatorios de Markov*

Como se mencionó en el capítulo anterior, la idea de los campos de Markov es integrar toda la información disponible de la lista otorgada de documentos recupera-

dos con la finalidad de generar un orden más apropiado para éstos. La hipótesis de esta estrategia es que por medio de considerar información que no es empleada en el proceso del re-ordenamiento propuesto (i.e., por medio de un documento virtual, y/o por medio de un resumen multi-documento) tal y como lo es la similitud entre los documentos recuperados, la posición inicialmente otorgada a los documentos y la información proporcionada por los *documentos ejemplo*, es posible distinguir de manera más precisa documentos relevantes de aquellos que no lo son, y de esta forma proporcionar un orden más adecuado para éstos.

Dado que para los campos aleatorios de Markov (MRF, por sus siglas en Inglés) se emplea la función de energía 4.2.8 expuesta en la sección 4.2.3, la cual involucra un factor $\varphi$ el cual determina la relevancia del potencial de interacción o del potencial de asociación, los experimentos realizados en esta sección muestran, además de los resultados variando los parámetros que hasta ahora se han expuesto, a los resultados obtenidos al variar este parámetro en la función de energía.

Para los experimentos realizados, a $\varphi$ se le dieron valores que van desde 1 hasta 0, indicando el primero que el valor de la función de energía es determinado únicamente por el potencial de interacción, mientras que el último indica entonces que el valor de la función de energía es determinado únicamente por el potencial de asociación. Al igual que en las secciones anteriores, $n$ se manejó con valores de 2, 5 y 10 *documentos ejemplo*, los cuales son obtenidos por medio de dos modalidades de retroalimentación: ciega y simulada.

### 5.6.1. Retroalimentación Ciega

En esta sección se muestran los resultados obtenidos con la estrategia propuesta para el re-ordenamiento empleando campos aleatorios de Markov considerando el caso de cuando los *documentos ejemplo* son obtenidos por medio de un proceso de retroalimentación ciega.

La tabla 5.31 muestra los resultados obtenidos cuando la forma de representación empleada para medir distancias entre los documentos (variables) dentro del MRF fue por medio de una representación *textual*.

Observe que en la mayoría de los casos es posible mejorar el desempeño del método base. Particularmente los mejores valores para la medida *R-prec* y *MAP* se dan cuando $n = 2$ y $\varphi = 0.7$, siendo estos resultados estadísticamente signifcativos. Note también

| *valor* $\varphi$ | *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|---|
| 1.0 | 2 *docs* | 0.14 | 0.27 | 0.24(-3.7) |
| | 5 *docs* | 0.16 | **0.28** | **0.25**(+0.9) |
| | 10 *docs* | 0.16 | **0.28** | 0.25(-0.8) |
| 0.7 | 2 *docs* | **0.17** | **0.29** | **0.26**(+4.6)* |
| | 5 *docs* | **0.17** | **0.28** | **0.26**(+3.8)* |
| | 10 *docs* | **0.17** | **0.28** | **0.26**(+3.9)* |
| 0.5 | 2 *docs* | **0.17** | **0.28** | **0.26**(+3.2) |
| | 5 *docs* | **0.17** | **0.28** | **0.26**(+3.1) |
| | 10 *docs* | **0.17** | **0.28** | **0.26**(+3.4) |
| 0.3 | 2 *docs* | **0.17** | **0.28** | **0.26**(+3) |
| | 5 *docs* | **0.17** | **0.28** | **0.26**(+3) |
| | 10 *docs* | **0.17** | **0.28** | **0.26**(+3.1) |
| 0.0 | 2 *docs* | **0.17** | **0.28** | **0.26**(+3) |
| | 5 *docs* | **0.17** | **0.28** | **0.26**(+3) |
| | 10 *docs* | **0.17** | **0.28** | **0.26**(+3) |
| *método base* | | *0.16* | *0.27* | *0.25* |

**Tabla 5.31: Experimento manejando: MRF; con representación: *BOW*; selección ciega**

que cuanto menor es el valor de $\varphi$ los resultados tienden a ser los mismos sin importar el valor de $n$, lo cual quiere decir que la información *a priori* o el valor del potencial de asociación por si sólo, empleando esta forma de representación no permite discriminar efectivamente entre elementos relevantes y no relevantes.

Las tablas 5.32, 5.33 y 5.34 muestran los resultados obtenidos cuando los documentos (variables) dentro del MRF fueron representados por medio de una forma *geográfica simple, geográfica expandida* y por medio de sus *coordenadas geográficas* respectivamente.

| *valor* $\varphi$ | *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|---|
| 1.0 | 2 *docs* | 0.15 | 0.27 | 0.25(-1.9) |
| | 5 *docs* | **0.16** | **0.28** | **0.26**(+2.1) |
| | 10 *docs* | 0.16 | 0.27 | **0.25**(+0.5) |
| 0.7 | 2 *docs* | **0.16** | **0.28** | **0.26**(+3.5)* |
| | 5 *docs* | **0.17** | **0.29** | **0.26**(+4.4)* |
| | 10 *docs* | 0.16 | **0.28** | **0.25**(+1.6) |
| 0.5 | 2 *docs* | **0.16** | **0.28** | **0.26**(+3.2) |
| | 5 *docs* | **0.17** | **0.29** | **0.26**(4.3)* |
| | 10 *docs* | **0.16** | **0.28** | **0.25**(1.8) |
| 0.3 | 2 *docs* | **0.16** | **0.28** | **0.26**(+3.1) |
| | 5 *docs* | **0.17** | **0.29** | **0.26**(4.2)* |
| | 10 *docs* | **0.17** | **0.28** | **0.26**(3.2) |
| 0.0 | 2 *docs* | 0.16 | **0.29** | **0.26**(+2.5) |
| | 5 *docs* | **0.17** | **0.29** | **0.26**(+2.8) |
| | 10 *docs* | **0.17** | **0.29** | **0.26**(3.1) |
| *método base* | | *0.16* | *0.27* | *0.25* |

**Tabla 5.32: Experimento manejando: MRF; con representación: *Geográfica Simple*; selección ciega**

Lo más importante a resaltar de estos experimentos es que con las tres formas de

| *valor* $\varphi$ | *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|---|
| 1.0 | 2 *docs* | 0.15 | 0.27 | 0.24(-2.5) |
| | 5 *docs* | 0.15 | 0.27 | 0.24(-4.7) |
| | 10 *docs* | 0.16 | **0.2813** | 0.2499(-0.1) |
| 0.7 | 2 *docs* | **0.16** | **0.28** | **0.26**(+2.6) |
| | 5 *docs* | **0.16** | **0.29** | **0.26**(+3.6)* |
| | 10 *docs* | 0.1618 | **0.28** | **0.26**(+1.9) |
| 0.5 | 2 *docs* | **0.16** | **0.28** | **0.25**(+1.9) |
| | 5 *docs* | **0.17** | **0.28** | **0.26**(+3.4) |
| | 10 *docs* | **0.16** | **0.28** | **0.26**(+2.6) |
| 0.3 | 2 *docs* | **0.16** | **0.28** | **0.25**(+1.4) |
| | 5 *docs* | **0.17** | **0.28** | **0.26**(+2.7) |
| | 10 *docs* | **0.17** | **0.28** | **0.26**(+2.8) |
| 0.0 | 2 *docs* | 0.16 | **0.28** | **0.25**(+1.2) |
| | 5 *docs* | **0.17** | **0.28** | **0.26**(+2.5) |
| | 10 *docs* | **0.17** | **0.28** | **0.26**(+2.7) |
| *método base* | | *0.16* | *0.27* | *0.25* |

**Tabla 5.33: Experimento manejando: MRF; con representación: *Geográfica Expandida*; selección ciega**

representación geográficas, en la mayoría de las configuraciones de $n$ y $\varphi$ es posible superar el desempeño del método base, hecho que no sucedió cuando se manejó un documento virtual y/o un resumen multi-documento. Este hecho quiere decir que el MRF tiene mayor éxito para discriminar entre elementos relevantes y no relevantes empleando únicamente la información geográfica contenida en los documentos, obteniendo un desempeño comparable con el obtenido al emplear toda la información disponible de los documentos (tabla 5.31). Sin embargo, no todos son resultados significativos, lo cual indica que los resultados no son necesariamente un efecto producido por la forma de representación.

| *valor* $\varphi$ | *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|---|
| 1.0 | 2 *docs* | 0.14 | 0.25 | 0.21(-16) |
| | 5 *docs* | 0.14 | 0.26 | 0.22(-12) |
| | 10 *docs* | 0.15 | 0.27 | 0.23(-6.2) |
| 0.7 | 2 *docs* | 0.15 | **0.28** | 0.24(-4.5) |
| | 5 *docs* | 0.15 | **0.28** | 0.24(-3.2) |
| | 10 *docs* | 0.16 | **0.28** | 0.25(-0.4) |
| 0.5 | 2 *docs* | **0.17** | **0.28** | **0.26**(+2.3) |
| | 5 *docs* | **0.17** | **0.28** | **0.26**(+2.2) |
| | 10 *docs* | **0.17** | **0.28** | **0.26**(+2.5) |
| 0.3 | 2 *docs* | **0.17** | **0.28** | **0.26**(+2.7) |
| | 5 *docs* | **0.17** | **0.28** | **0.26**(+2.6) |
| | 10 *docs* | **0.17** | **0.28** | **0.26**(+2.8) |
| 0.0 | 2 *docs* | **0.17** | **0.28** | **0.26**(+2.6) |
| | 5 *docs* | **0.17** | **0.28** | **0.26**(+2.6) |
| | 10 *docs* | **0.17** | **0.28** | **0.26**(+2.7) |
| *método base* | | *0.16* | *0.27* | *0.25* |

**Tabla 5.34: Experimento manejando: MRF; con representación: *Coordenadas Geográficas*; selección ciega**

Además, se sigue cumpliendo que los resultados empleando una forma de representación a través de los términos geográficos expandidos (tabla 5.33), son en general menores a los obtenidos con la representación *geográfica simple* (tabla 5.32). Y de igual manera, para el caso de una forma de representación con *coordenadas geográficas* (tabla 5.34), los resultados obtenidos en general son los más bajos de estos tres tipos de representación geográfica y la única donde ningún experimento logro ser significativo.

Las tablas siguientes muestran los resultados obtenidos al momento de combinar las fuentes de información temática y geográfica de los documentos por medio de la fórmula 4.3.1 descrita en la sección 4.3. La tabla 5.35 muestra los resultados obtenidos cuando la parte geográfica de los documentos es representada por medio de las entidades geográficas simples. La tabla 5.36 muestra los resultados cuando la parte geográfica se representó por medio de las entidades geográficas expandidas, y finalmente la tabla 5.37 muestra el desempeño del sistema cuando la parte geográfica se representó por medio de coordenadas geográficas.

En estos experimentos (tablas 5.35, 5.36 y 5.37) se supera en la mayoría de los casos al método base, de ahí que se remarcan en negritas por cada columna sólo las configuraciones que obtuvieron los valores más altos.

| *valor* | *num* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\varphi$ | *docs* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| 1.0 | 2 *docs* | 0.14 | 0.26 | 0.24(-3.9) | 0.14 | 0.26 | 0.23(-4.5) | 0.13 | 0.24 | 0.22(-12) | 0.14 | 0.26 | 0.23(-9) |
| | 5 *docs* | 0.16 | **0.28** | 0.26(+2) | 0.16 | 0.28 | 0.25(+0.5) | 0.15 | 0.27 | 0.24(-3.5) | 0.15 | 0.28 | 0.24(-2.7) |
| | 10 *docs* | 0.16 | 0.27 | 0.25(+1.2) | 0.16 | 0.27 | 0.25(+1) | 0.15 | 0.27 | 0.25(-2) | 0.16 | 0.28 | 0.25(-1.2) |
| 0.7 | 2 *docs* | 0.17 | 0.28 | 0.26(+4)* | 0.16 | 0.28 | **0.26**(+4.4)* | **0.17** | **0.29** | **0.26**(+5.4)* | 0.16 | **0.29** | **0.26**(+4.9)* |
| | 5 *docs* | **0.17** | **0.29** | 0.26(+5)* | 0.17 | **0.28** | 0.26(+4.3)* | 0.17 | 0.28 | 0.26(+3.2) | **0.17** | 0.28 | 0.26(+4)* |
| | 10 *docs* | 0.16 | 0.28 | 0.26(+2.3) | 0.16 | 0.28 | 0.26(+2.4) | 0.17 | 0.29 | 0.26(+3.2) | 0.17 | 0.28 | 0.26(+4)* |
| 0.5 | 2 *docs* | 0.17 | 0.28 | **0.26**(+4.4)* | **0.17** | 0.28 | 0.26(+4.2)* | 0.17 | 0.28 | 0.26(+3.7)* | 0.17 | 0.28 | 0.26(+3.3) |
| | 5 *docs* | 0.17 | 0.28 | 0.26(+4.2)* | 0.17 | 0.28 | 0.26(+3.6)* | 0.17 | 0.28 | 0.26(+3.4) | 0.17 | 0.28 | 0.26(+3.2) |
| | 10 *docs* | 0.16 | 0.28 | 0.26(+2.5) | 0.17 | 0.28 | 0.26(+3.4) | 0.17 | 0.28 | 0.26(+3.6)* | 0.17 | 0.28 | 0.26(+3.4) |
| 0.3 | 2 *docs* | 0.17 | 0.28 | 0.26(+3.5)* | 0.17 | 0.28 | 0.26(+3.3) | 0.17 | 0.28 | 0.26(+3.1) | 0.17 | 0.28 | 0.26(+3.1) |
| | 5 *docs* | 0.17 | 0.28 | 0.26(+3.6)* | 0.17 | 0.28 | 0.26(+3.1) | 0.17 | 0.28 | 0.26(+3) | 0.17 | 0.28 | 0.26(+3) |
| | 10 *docs* | 0.17 | 0.28 | 0.26(+3.4) | 0.17 | 0.28 | 0.26(+3.2) | 0.17 | 0.28 | 0.26(+3.2) | 0.17 | 0.28 | 0.26(+3.1) |
| 0.0 | 2 *docs* | 0.17 | 0.28 | 0.26(+3.2) | 0.17 | 0.28 | 0.26(+3) | 0.17 | 0.28 | 0.26(+3) | 0.17 | 0.28 | 0.26(+3) |
| | 5 *docs* | 0.17 | 0.28 | 0.26(+3.1) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) |
| | 10 *docs* | 0.17 | 0.28 | 0.26(+3.1) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+3) |
| *método base: p@100 = 0.16 ; R-prec = 0.27 ; MAP = 0.25* | | | | | | | | | | | | | |

**Tabla 5.35: Experimento manejando: MRF; con representación: *Temática y Geográfica Simple*; selección ciega**

Como factor común en estos experimentos se puede observar que cuando $\varphi$ oscila entre valores de 0.7 y 0.5 es cuando se logran obtener los mejores resultados. Además de esto, tal y como venía sucediendo en experimentos previos, cuando $\lambda$=0.6 es cuando

| *valor* | *num* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\varphi$ | *docs* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| 1.0 | 2 *docs* | 0.14 | 0.26 | 0.23(-7) | 0.14 | 0.27 | 0.24(-5) | 0.14 | 0.26 | 0.23(-8.1) | 0.15 | 0.26 | 0.22(-8.8) |
| | 5 *docs* | 0.14 | 0.26 | 0.24(-5.8) | 0.15 | 0.27 | 0.25(-1.8) | 0.14 | 0.27 | 0.24(-5.4) | 0.15 | 0.27 | 0.24(-5.1) |
| | 10 *docs* | 0.15 | 0.27 | 0.24(-2.3) | 0.15 | 0.28 | 0.25(-1) | 0.15 | 0.27 | 0.24(-3) | 0.14 | 0.26 | 0.24(-5.6) |
| 0.7 | 2 *docs* | 0.16 | 0.28 | 0.26(+3.4) | 0.17 | 0.28 | 0.26(+3.6)* | **0.17** | **0.29** | **0.26**(+5)* | 0.17 | **0.29** | **0.26**(+5)* |
| | 5 *docs* | 0.17 | **0.29** | **0.26**(+3.8)* | 0.17 | **0.29** | **0.26**(+4.2)* | 0.17 | 0.28 | 0.26(+4.2)* | 0.17 | 0.28 | 0.26(+4)* |
| | 10 *docs* | 0.16 | 0.28 | 0.26(+1.9) | 0.16 | 0.28 | 0.26(+2.6) | 0.17 | 0.28 | 0.26(+3.7)* | **0.17** | 0.29 | 0.26(+4)* |
| 0.5 | 2 *docs* | **0.17** | 0.28 | 0.26(+3.6)* | **0.17** | 0.28 | 0.26(+4)* | 0.17 | 0.28 | 0.26(+3.7)* | 0.17 | 0.28 | 0.26(+3.2) |
| | 5 *docs* | 0.17 | 0.28 | 0.26(+3.7)* | 0.17 | 0.28 | 0.26(+3.7)* | 0.17 | 0.28 | 0.26(+3.4) | 0.17 | 0.28 | 0.26(+3.1) |
| | 10 *docs* | 0.17 | 0.28 | 0.26(+3.4) | 0.17 | 0.28 | 0.26(+3.7)* | 0.17 | 0.28 | 0.26(+3.6)* | 0.17 | 0.28 | 0.26(+3.4) |
| 0.3 | 2 *docs* | 0.17 | 0.28 | 0.26(+3.6)* | 0.17 | 0.28 | 0.26(+3.2) | 0.17 | 0.28 | 0.26(+3) | 0.17 | 0.28 | 0.26(+3) |
| | 5 *docs* | 0.17 | 0.28 | 0.26(+3.3) | 0.17 | 0.28 | 0.26(+3.1) | 0.17 | 0.28 | 0.26(+3) | 0.17 | 0.28 | 0.26(+3) |
| | 10 *docs* | 0.17 | 0.28 | 0.26(+3.2) | 0.17 | 0.28 | 0.26(+3) | 0.17 | 0.28 | 0.26(+3) | 0.17 | 0.28 | 0.26(+3) |
| 0.0 | 2 *docs* | 0.17 | 0.28 | 0.26(+3) | 0.17 | 0.28 | 0.26(+3) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) |
| | 5 *docs* | 0.17 | 0.28 | 0.26(+3) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) |
| | 10 *docs* | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) |
| *método base: p@100 = 0.16 ; R-prec = 0.27 ; MAP = 0.25* | | | | | | | | | | | | | |

**Tabla 5.36: Experimento manejando: MRF; con representación: *Temática y Geográfica Expandida*; selección ciega**

| *valor* | *num* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\varphi$ | *docs* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| 1.0 | 2 *docs* | 0.13 | 0.24 | 0.20(-20) | 0.13 | 0.24 | 0.20(-20) | 0.13 | 0.25 | 0.21(-16) | 0.14 | 0.25 | 0.22(-13) |
| | 5 *docs* | 0.14 | 0.26 | 0.22(-12) | 0.14 | 0.26 | 0.22(-12) | 0.14 | 0.26 | 0.22(-12) | 0.15 | 0.27 | 0.23(-7) |
| | 10 *docs* | 0.15 | 0.27 | 0.23(-8) | 0.15 | 0.27 | 0.23(-8) | 0.15 | 0.27 | 0.23(-8) | 0.15 | 0.27 | 0.24(-4) |
| 0.7 | 2 *docs* | 0.16 | 0.27 | 0.24(-3) | 0.16 | 0.28 | 0.25(-1.2) | 0.17 | 0.28 | 0.26(+2.7) | 0.17 | 0.28 | 0.26(+3.3) |
| | 5 *docs* | 0.16 | 0.28 | 0.25(-1.6) | 0.16 | 0.28 | 0.25(-0.1) | 0.17 | **0.28** | **0.26**(+3) | **0.17** | **0.28** | 0.26(+3.4) |
| | 10 *docs* | 0.16 | **0.28** | 0.25(+0.6) | 0.16 | **0.28** | 0.25(+1) | 0.17 | 0.28 | 0.26(+3) | **0.17** | 0.28 | **0.26***(+3.6) |
| 0.5 | 2 *docs* | **0.17** | 0.28 | 0.26(+2.6) | **0.17** | 0.28 | **0.26** (+3) | **0.17** | 0.28 | **0.26**(+3) | 0.17 | 0.28 | 0.26(+2.9) |
| | 5 *docs* | 0.17 | 0.28 | 0.26(+2.7) | 0.17 | 0.28 | **0.26**(+3) | 0.17 | 0.28 | 0.26(+3) | 0.17 | 0.28 | 0.26(+2.9) |
| | 10 *docs* | 0.17 | 0.28 | 0.26(+2.8) | 0.17 | 0.28 | **0.26** (+3) | 0.17 | 0.28 | **0.26**(+2.9) | 0.17 | 0.28 | 0.26(+3) |
| 0.3 | 2 *docs* | 0.17 | 0.28 | **0.26**(+2.8) | 0.17 | 0.28 | 0.26(+3) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) |
| | 5 *docs* | 0.17 | 0.28 | **0.26**(+2.8) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) |
| | 10 *docs* | 0.17 | 0.28 | **0.26**(+2.8) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) |
| 0.0 | 2 *docs* | 0.17 | 0.28 | 0.26(+2.8) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) |
| | 5 *docs* | 0.17 | 0.28 | 0.26(+2.8) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) |
| | 10 *docs* | 0.17 | 0.28 | 0.26(+2.8) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) | 0.17 | 0.28 | 0.26(+2.9) |
| *método base: p@100 = 0.16 ; R-prec = 0.27 ; MAP = 0.25* | | | | | | | | | | | | | |

**Tabla 5.37: Experimento manejando: MRF; con representación: *Temática y Coordenadas Geográficas*; selección ciega**

se obtiene el mejor desempeño tanto para el caso de la tabla 5.35 como para la tabla 5.36, que corresponden a cuando la parte geográfica se representó por medio de entidades geográficas simples y entidades geográficas expandidas respectivamente. Nótese que el mejor resultado obtenido hasta el momento se da en la tabla 5.35, con $\lambda$=0.6, $\varphi$=0.7 y $n = 2$.

Otro factor repetitivo que se observa en estos experimentos es que cuando la parte geográfica es representada por medio de coordenadas geográficas, el mejor resultado se da cuando $\lambda$=0.8, i.e., sigue siendo la forma de representación que aporta menor información discriminativa.

### 5.6.2. Retroalimentación Simulada

Los experimentos realizados en esta sección consideran un proceso de retroalimentación de relevancia simulada. De la misma manera, como se ha venido presentando en secciones anteriores, los resultados marcados en letras negritas equivalen a las configuraciones que fueron capaces de superar al simple proceso de retroalimentación (Tabla 5.9).

La tabla 5.38 muestra los resultados obtenidos al manejar una forma de representación *BOW* para los documentos dentro del MRF. Como es posible observar, aunque hay varias configuraciones que superan al proceso de retroalimentación (principalmente en la medida MAP), la mayoría de estos resultados no superan de manera notable al proceso de retroalimentación, cosa que no sucede cuando manejamos a los *documentos ejemplo* por medio de un documento virtual y/o por medio de un resumen multi-documento. Esto es posible comprobarlo por medio de observar los valores de ganancia del método, pues para el caso de los MRF estos valores oscilan entre un 3 %, mientras que para el documento virtual y/o el resumen multi-documento esta ganancia va desde un 11 % hasta un 25 % aproximadamente.

Las tablas siguientes (Tabla 5.39, 5.40 y 5.41) muestran los resultados de la estrategia de los MRF tras representar a los documentos por medio de sólo *entidades geográficas simples*, *entidades geográficas expandidas* y por medio de *coordenadas geográficas* respectivamente.

Como es posible observar, en muchos casos de estas tres formas distintas de representación geográfica es posible superar el método base y también al proceso de retroalimentación. Note también que los mejores resultados en general se dan cuando

| *valor* $\varphi$ | *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|---|
| 1.0 | 2 *docs* | 0.16 | **0.30** | 0.30(+3.8) |
| | 5 *docs* | **017** | **0.38** | **0.39**(+4.4) |
| | 10 *docs* | **0.18** | **0.49** | **0.51**(+1.2) |
| 0.7 | 2 *docs* | **0.17** | **0.30** | **0.30**(+5) |
| | 5 *docs* | **0.17** | **0.37** | **0.38**(+2.1) |
| | 10 *docs* | **0.18** | **0.48** | **0.51**(+0.7) |
| 0.5 | 2 *docs* | 0.17 | 0.30 | **0.30**(+4.1) |
| | 5 *docs* | 0.17 | 0.36 | **0.38**(+1.6) |
| | 10 *docs* | **0.18** | **0.48** | **0.50**(+0.4) |
| 0.3 | 2 *docs* | 0.17 | 0.30 | **0.30**(+3.9) |
| | 5 *docs* | 0.17 | 0.36 | **0.38**(+1.5) |
| | 10 *docs* | 0.18 | 0.48 | **0.50**(+0.3) |
| 0.0 | 2 *docs* | 0.17 | 0.30 | **0.30**(+3.9) |
| | 5 *docs* | 0.17 | 0.36 | **0.38**(+1.4) |
| | 10 *docs* | 0.18 | 0.48 | **0.50**(+0.2) |

**Tabla 5.38: Experimento manejando: MRF; con representación: *BOW*; selección simulada**

| *valor* $\varphi$ | *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|---|
| 1.0 | 2 *docs* | **0.17** | **0.31** | **0.31**(+6.1) |
| | 5 *docs* | **0.17** | **0.37** | **0.38**(+3.1) |
| | 10 *docs* | **0.19** | **0.49** | **0.51**(+2.3)* |
| 0.7 | 2 *docs* | **0.18** | **0.32** | **0.31**(+8.7)* |
| | 5 *docs* | **0.18** | **0.37** | **0.39**(+4.3) |
| | 10 *docs* | **0.19** | **0.49** | **0.51**(+2.1)* |
| 0.5 | 2 *docs* | **0.18** | **0.31** | **0.31**(+7.8)* |
| | 5 *docs* | **0.18** | **0.37** | **0.39**(+3.8) |
| | 10 *docs* | **0.18** | **0.49** | **0.51**(+1.7) |
| 0.3 | 2 *docs* | **0.17** | **0.31** | **0.31**(+7.2) |
| | 5 *docs* | **0.18** | **0.37** | **0.39**(+3.5) |
| | 10 *docs* | **0.19** | **0.49** | **0.51**(+1.5) |
| 0.0 | 2 *docs* | **0.17** | **0.32** | **0.31**(+7.4) |
| | 5 *docs* | **0.17** | **0.37** | **0.38**(+2.7) |
| | 10 *docs* | **0.18** | **0.49** | **0.51**(+1.1) |

**Tabla 5.39: Experimento manejando: MRF; con representación: *Geográfica Simple*; selección simulada**

| *valor* $\varphi$ | *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|---|
| 1.0 | 2 *docs* | 0.15 | 0.29 | 0.29(+1) |
| | 5 *docs* | 0.16 | 0.36 | 0.37(-1.4) |
| | 10 *docs* | **0.18** | **0.49** | **0.51**(+1.2) |
| 0.7 | 2 *docs* | **0.17** | **0.31** | **0.31**(+6.7) |
| | 5 *docs* | **0.17** | **0.37** | **0.39**(+3.6) |
| | 10 *docs* | **0.18** | **0.48** | **0.51**(+1.4) |
| 0.5 | 2 *docs* | **0.17** | **0.30** | **0.31**(+5.6) |
| | 5 *docs* | **0.17** | **0.35** | **0.38**(+2.7) |
| | 10 *docs* | **0.18** | **0.48** | **0.51**(+0.9) |
| 0.3 | 2 *docs* | **0.17** | **0.30** | **0.30**(+4.6) |
| | 5 *docs* | **0.17** | **0.37** | **0.38**(+2.1) |
| | 10 *docs* | **0.18** | **0.48** | **0.51**(+0.8) |
| 0.0 | 2 *docs* | **0.17** | **0.30** | **0.30**(+4.2) |
| | 5 *docs* | **0.17** | **0.36** | **0.38**(+1.6) |
| | 10 *docs* | **0.18** | **0.48** | **0.50**(+0.4) |

**Tabla 5.40: Experimento manejando: MRF; con representación: *Geográfica Expandida*; selección simulada**

se emplea sólo a las entidades geográficas simples para representar a los documentos (Tabla 5.39), siendo éstos comparables y en general mejores que los resultados obtenidos al utilizar una forma de representación *BOW* (Tabla 5.38). Este hecho soporta la observación hecha en la sección anterior, donde se dijo que los MRF tienen mayor éxito para discriminar entre elementos relevantes y no relevantes empleando únicamente la información geográfica contenida en los documentos.

| *valor* $\varphi$ | *num docs* | *p@100* | *R-prec* | *MAP* |
|---|---|---|---|---|
| 1.0 | 2 *docs* | 0.14 | 0.27 | 0.25(-12) |
| | 5 *docs* | 0.16 | 0.35 | 0.36(4.3) |
| | 10 *docs* | 0.18 | **0.48** | 0.49(-1.5) |
| 0.7 | 2 *docs* | 0.16 | 0.30 | 0.29(-1.1) |
| | 5 *docs* | 0.17 | 0.37 | 0.38(+0.4) |
| | 10 *docs* | **0.18** | 0.49 | **0.50**(+0.3) |
| 0.5 | 2 *docs* | **0.17** | 0.30 | 0.30(+3.8) |
| | 5 *docs* | **0.17** | 0.36 | **0.38**(+1.4) |
| | 10 *docs* | **0.18** | **0.48** | **0.50**(+0.3) |
| 0.3 | 2 *docs* | 0.17 | 0.30 | 0.30(+3.7) |
| | 5 *docs* | 0.17 | 0.36 | 0.37(+1.3) |
| | 10 *docs* | **0.18** | **0.48** | **0.50**(+0.2) |
| 0.0 | 2 *docs* | 0.17 | 0.30 | 0.30(+3.6) |
| | 5 *docs* | 0.17 | 0.36 | 0.38(+1.2) |
| | 10 *docs* | **0.18** | **0.48** | **0.50**(+0.1) |

**Tabla 5.41: Experimento manejando: MRF; con representación: *Coordenadas Geográficas*; selección simulada**

Es importante también en este punto hacer notar que los resultados que logran ser mejores, no superan por mucho al proceso único de la retroalimentación de relevancia (Vea tabla 5.9). Esto indica que el MRF aunque otorga en ocasiones resultados significativos, se está quedando en máximos locales, pues como hemos visto en secciones anteriores, el proceso de retroalimentación simulada es superado notablemente tanto por la estrategia del re-ordenamiento a través de un documento virtual como por la que propone hacerlo por medio de un resumen multi-documento.

Las siguientes tablas muestran los resultados obtenidos al representar los documentos por medio de la combinación de fuentes de información, i.e., la información temática y la información geográfica combinadas por medio de la fórmula 4.3.1 expuesta en la sección 4.3. La tabla 5.42 muestra los resultados obtenidos cuando la parte geográfica involucra sólo a las entidades geográficas simples. Por otro lado, la tabla 5.43 muestra los resultados obtenidos cuando la parte geográfica se compone de las entidades geográficas expandidas, y finalmente, la tabla 5.44 muestra los resultados cuando lo geográfico se representa por medio de coordenadas geográficas.

En estas tablas se marcan en negritas los mejores resultados alcanzados en cada

columna así como para los distintos valores de $n$.

| *valor* | *num* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\varphi$ | *docs* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| 1.0 | 2 *docs* | 0.17 | **0.31** | 0.30(+5.5) | 0.16 | 0.30 | 0.30(+4) | 0.16 | 0.29 | 0.29(-0.9) | 0.16 | 0.30 | 0.29(-0.7) |
| | 5 *docs* | 0.17 | **0.37** | **0.39**(+3.5) | **0.18** | **0.37** | **0.39**(+3.5) | **0.17** | 0.37 | 0.39(+1.3) | 0.17 | **0.37** | **0.38**(+2.8) |
| | 10 *docs* | **0.19** | **0.49** | **0.51**(+1.9) | **0.19** | **0.49** | **0.51**(+2.2)* | **0.19** | **0.49** | **0.51**(+1.6) | 0.18 | **0.49** | **0.51**(+1) |
| 0.7 | 2 *docs* | **0.18** | 0.31 | **0.31**(+7.7)* | **0.17** | **0.31** | **0.31**(+6.6) | **0.17** | **0.30** | **0.31**(+6) | **0.17** | **0.30** | **0.30**(+5.3) |
| | 5 *docs* | **0.18** | **0.37** | 0.39(+3.5) | 0.17 | 0.37 | 0.38(+2.9) | 0.17 | **0.37** | **0.38**(+2.7) | **0.17** | 0.37 | 0.38(+2.3) |
| | 10 *docs* | 0.18 | 0.49 | 0.51(+1.6) | 0.18 | 0.48 | 0.51(+1.4) | 0.18 | 0.48 | 0.51(+1.1) | **0.18** | 0.48 | 0.51(+0.9) |
| 0.5 | 2 *docs* | 0.17 | 0.30 | 0.31(+6) | 0.17 | 0.30 | 0.30(+4.9) | 0.17 | 0.30 | 0.30(+4.3) | 0.17 | 0.30 | 0.30(+4.1) |
| | 5 *docs* | 0.17 | 0.37 | 0.38(+2.8) | 0.17 | 0.36 | 0.38(+2) | 0.17 | 0.36 | 0.38(+1.7) | 0.17 | 0.36 | 0.37(+1.6) |
| | 10 *docs* | 0.18 | 0.48 | 0.51(+1) | 0.18 | 0.48 | 0.50(+0.6) | 0.18 | 0.48 | 0.50(+0.5) | 0.18 | 0.48 | 0.50(+0.4) |
| 0.3 | 2 *docs* | 0.17 | 0.30 | 0.30(+4.8) | 0.17 | 0.30 | 0.30(+4.1) | 0.17 | 0.30 | 0.30(+4) | 0.17 | 0.30 | 0.30(+4) |
| | 5 *docs* | 0.17 | 0.36 | 0.38(+2) | 0.17 | 0.36 | 0.38(+1.6) | 0.17 | 0.36 | 0.38(+1.5) | 0.17 | 0.36 | 0.38(+1.5) |
| | 10 *docs* | 0.18 | 0.48 | 0.50(+0.6) | 0.18 | 0.48 | 0.50(+0.3) | 0.18 | 0.48 | 0.50(+0.3) | 0.18 | 0.48 | 0.50(+0.3) |
| 0.0 | 2 *docs* | 0.17 | 0.30 | 0.30(+4) | 0.17 | 0.30 | 0.30(+4) | 0.17 | 0.30 | 0.30(+4) | 0.17 | 0.30 | 0.30(+3.9) |
| | 5 *docs* | 0.17 | 0.36 | 0.38(+1.5) | 0.17 | 0.36 | 0.38(+1.4) | 0.17 | 0.36 | 0.38(+1.4) | 0.17 | 0.36 | 0.38(+1.4) |
| | 10 *docs* | 0.18 | 0.48 | 0.50(+0.3) | 0.18 | 0.48 | 0.50(+0.2) | 0.18 | 0.48 | 0.50(+0.2) | 0.18 | 0.48 | 0.50(+0.2) |

**Tabla 5.42: Experimento manejando: MRF; con representación: *Temática y Geográfica Simple*; selección simulada**

Entre las cosas importantes a resaltar en estos experimentos, está el hecho de que en los tres experimentos (tablas 5.42, 5.43 y 5.44 ) se observa que a valores pequeños de $\lambda$ (0.2 y 0.4) se obtienen los mejores resultados en las diferentes configuraciones. Lo cual si recordamos, quiere decir que se le está dando mayor importancia a la parte geográfica de los documentos, y como hemos observado que los MRF se desempeñan bien con la información geográfica por si sola, podemos concluir que este comportamiento se debe principalmente a que en efecto la información geográfica le permite diferenciar de mejor manera al MRF entre elementos relevantes y no relevantes.

Otro detalle importante en estos experimentos es el hecho de que los mejores resultados se obtienen en general cuando $\varphi$ tiene un valor de 1.0 y 0.7, lo cual indica que, para el caso de la retroalimentación simulada, la información *a priori* que se le da al MRF no aporta suficiente ganancia, en su lugar el potencial de interacción es el que ayuda a determinar de mejor forma el valor relevante o no relevante de los documentos.

Finalmente, es importante hacer notar que al igual que en los experimentos previos, la ganancia con respecto al proceso único de retroalimentación de relevancia (Vea tabla 5.9) en general no es significativa, lo que sugiere que el algoritmo de optimización ICM se está quedando en máximos locales.

| *valor* | *num* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\varphi$ | *docs* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| | 2 *docs* | 0.14 | 0.29 | 0.28(-1.4) | 0.14 | 0.29 | 0.29(-0.3) | 0.15 | 0.29 | 0.28(-2) | 0.16 | 0.30 | 0.29(-1.1) |
| 1.0 | 5 *docs* | 0.16 | 0.36 | 0.37(-1.9) | 0.17 | **0.37** | 0.38(+0.9) | 0.16 | 0.37 | 0.37(+0.2) | 0.17 | **0.37** | 0.38(+1.6) |
| | 10 *docs* | 0.18 | 0.48 | 0.51(+0.9) | 0.18 | 0.49 | 0.51(+1.2) | **0.18** | **0.49** | **0.51**(+1) | 0.18 | **0.49** | **0.51**(+1.3) |
| | 2 *docs* | **0.17** | **0.31** | **0.31**(+6.9) | **0.17** | **0.31** | **0.30**(+6.6) | **0.17** | **0.30** | **0.30**(+5.5) | **0.17** | **0.30** | **0.30**(+5.3) |
| 0.7 | 5 *docs* | **0.18** | **0.37** | **0.39**(+3.5) | **0.17** | 0.37 | **0.38**(+2.9) | **0.17** | **0.37** | **0.38**(+2.5) | **0.17** | 0.37 | **0.38**(+2.3) |
| | 10 *docs* | **0.18** | **0.49** | **0.51**(+1.2) | **0.18** | **0.49** | **0.51**(+1.3) | 0.18 | 0.48 | 0.51(+0.94) | **0.18** | 0.48 | 0.51(+0.9) |
| | 2 *docs* | **0.17** | 0.30 | 0.31(+5.9) | 0.17 | 0.30 | 0.30(+4.6) | 0.17 | 0.30 | 0.30(+4.3) | 0.17 | 0.30 | 0.30(+4) |
| 0.5 | 5 *docs* | 0.17 | 0.37 | 0.38(+2.5) | 0.17 | 0.36 | 0.38(+1.9) | 0.17 | 0.36 | 0.38(+1.8) | 0.17 | 0.36 | 0.38(+1.6) |
| | 10 *docs* | 0.18 | 0.48 | 0.50(+0.9) | 0.18 | 0.48 | 0.50(+0.6) | 0.18 | 0.48 | 0.50(+0.5) | 0.18 | 0.4800 | 0.50(+0.4) |
| | 2 *docs* | 0.17 | 0.30 | 0.30(+4.3) | 0.17 | 0.30 | 0.30(+4.1) | 0.17 | 0.30 | 0.30(+3.9) | 0.17 | 0.30 | 0.30(+3.9) |
| 0.3 | 5 *docs* | 0.17 | 0.36 | 0.38(+1.7) | 0.17 | 0.36 | 0.38(+1.5) | 0.17 | 0.36 | 0.38(+1.5) | 0.17 | 0.36 | 0.38(+1.5) |
| | 10 *docs* | 0.18 | 0.48 | 0.50(+0.4) | 0.18 | 0.48 | 0.50(+0.3) | 0.18 | 0.48 | 0.50(+0.3) | 0.18 | 0.48 | 0.50(+0.3) |
| | 2 *docs* | 0.17 | 0.30 | 0.30(+3.9) | 0.17 | 0.30 | 0.30(+3.9) | 0.17 | 0.30 | 0.30(+3.9) | 0.17 | 0.30 | 0.30(+3.9) |
| 0.0 | 5 *docs* | 0.17 | 0.36 | 0.38(+1.4) | 0.17 | 0.36 | 0.38(+1.4) | 0.17 | 0.36 | 0.38(+1.4) | 0.17 | 0.36 | 0.38(+1.4) |
| | 10 *docs* | 0.18 | 0.48 | 0.50(+0.2) | 0.18 | 0.48 | 0.50(+0.2) | 0.18 | 0.48 | 0.50(+0.2) | 0.18 | 0.48 | 0.50(+0.2) |

**Tabla 5.43: Experimento manejando: MRF; con representación: *Temática y Geográfica Expandida*; selección simulada**

| *valor* | *num* | $\lambda = 0{,}2$ | | | $\lambda = 0{,}4$ | | | $\lambda = 0{,}6$ | | | $\lambda = 0{,}8$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\varphi$ | *docs* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* | *p@100* | *R-prec* | *MAP* |
| | 2 *docs* | 0.14 | 0.26 | 0.25(-14) | 0.14 | 0.27 | 0.25(-14) | 0.14 | 0.28 | 0.26(-1) | 0.14 | 0.28 | 0.27(-8.1) |
| 1.0 | 5 *docs* | 0.16 | 0.35 | 0.36(-4.7) | 0.16 | 0.36 | 0.36(-3.6) | 0.16 | 0.36 | 0.36(-3.3) | 0.16 | 0.36 | 0.37(-1.9) |
| | 10 *docs* | 0.18 | 0.48 | 0.50(-1) | 0.18 | 0.48 | 0.50(-1) | 0.18 | 0.49 | 0.50(-0.3) | **0.19** | **0.49** | **0.50**(+0.8) |
| | 2 *docs* | 0.16 | **0.30** | 0.29(+0.6) | 0.17 | **0.30** | 0.30(+2.5) | **0.17** | **0.30** | **0.30**(+4) | 0.17 | 0.30 | **0.30**(+4.2) |
| 0.7 | 5 *docs* | 0.17 | **0.37** | 0.38(+0.7) | **0.17** | **0.37** | 0.38(+1.3) | **0.17** | **0.37** | **0.38**(+2) | **0.17** | **0.36** | **0.38**(+1.8) |
| | 10 *docs* | **0.18** | **0.49** | **0.50**(+0.6) | **0.18** | **0.48** | **0.50**(+0.7) | **0.18** | **0.48** | **0.51**(+0.6) | 0.18 | 0.48 | 0.50(+0.5) |
| | 2 *docs* | **0.17** | 0.30 | **0.30**(+3.9) | 0.17 | 0.30 | **0.30**(+4) | 0.17 | 0.30 | 0.30(+4) | 0.17 | 0.30 | 0.30(+4) |
| 0.5 | 5 *docs* | 0.17 | 0.36 | **0.38**(+1.4) | 0.17 | 0.36 | **0.38**(+1.4) | 0.17 | 0.36 | 0.38(+1.4) | 0.17 | 0.36 | 0.38(+1.4) |
| | 10 *docs* | 0.18 | 0.48 | 0.50(+0.3) | 0.18 | 0.48 | 0.50(+0.3) | 0.18 | 0.48 | 0.50(+0.3) | 0.18 | 0.48 | 0.50(+0.2) |
| | 2 *docs* | 0.17 | 0.30 | 0.30(+3.8) | 0.17 | 0.30 | 0.30(+3.9) | 0.17 | 0.30 | 0.30(+3.8) | 0.17 | 0.30 | 0.30(+3.8) |
| 0.3 | 5 *docs* | 0.17 | 0.36 | 0.38(+1.3) | 0.17 | 0.36 | 0.38(+1.3) | 0.17 | 0.36 | 0.38(+1.3) | 0.17 | 0.36 | 0.38(+1.4) |
| | 10 *docs* | 0.18 | 0.48 | 0.50(+0.2) | 0.18 | 0.48 | 0.50(+0.2) | 0.18 | 0.48 | 0.50(+0.2) | 0.18 | 0.48 | 0.50(+0.2) |
| | 2 *docs* | 0.17 | 0.30 | 0.30(+3.8) | 0.17 | 0.30 | 0.30(+3.8) | 0.17 | 0.30 | 0.30(+3.8) | 0.17 | 0.30 | 0.30(+3.8) |
| 0.0 | 5 *docs* | 0.17 | 0.36 | 0.38(+1.3) | 0.17 | 0.36 | 0.38(+1.3) | 0.17 | 0.36 | 0.38(+1.3) | 0.17 | 0.36 | 0.38(+1.3) |
| | 10 *docs* | 0.18 | 0.48 | 0.50(+0.1) | 0.18 | 0.48 | 0.50(+0.2) | 0.18 | 0.48 | 0.50(+0.2) | 0.18 | 0.48 | 0.50(+0.2) |

**Tabla 5.44: Experimento manejando: MRF; con representación: *Temática y Coordenadas Geográficas*; selección simulada**

## 5.7. Discusión

En este capítulo se presentó el conjunto completo de experimentos realizados para evaluar la pertinencia del método de re-ordenamiento propuesto. Tres formas distintas de manejar la información de los *documentos ejemplo* fueron propuestas y evaluadas, *i)* a través de un documento virtual, *ii)* por medio de un resumen multi-documento, y finalmente, *iii)* por medio de integrar toda la información disponible en un MRF. Estas tres formas de manejar la información de los *documentos ejemplo* fue combinada con diferentes formas de representación de los documentos que van desde una representación de tipo bolsa de palabras o *BOW* hasta una forma combinada de representaciones temática y geográfica. Además de esto, una tercer variante que se incluyó en los experimentos fue la forma de selección de los *documentos ejemplo*, las cuáles fueron de manera ciega y simulada.

La figura 5.1 muestra de manera resumida los mejores resultados obtenidos con las diferentes propuestas de solución bajo un esquema de selección ciega de los *documentos ejemplo*. Las columnas con el símbolo (*) denotan aquellos resultados que logran invalidar la hipótesis nula con una confianza de al menos un 90 %. Además, sobre cada columna se muestra el porcentaje de ganancia (+) o pérdida (-) de cada experimento con respecto al método base. Todos los experimentos mostrados en esta figura corresponden a una forma de representación *temática* y *geográfica simple*.

Note que para el caso cuando los *documentos ejemplo* son manejados por medio de un documento virtual (DocVirt) y un resumen multi-documento (Resumen multi-doc) es necesario considerar mínimo 10 *documentos ejemplo* para lograr obtener resultados significativos. De manera particular para el caso cuando se manejo un resumen multi-documento, los resultados son en general bajos, lo que supone la posibilidad de que el resumen que se está construyendo sea demasiado pequeño para completar la tarea con éxito, pues recordemos que apenas conservamos un 40 % de la información original. Sin embargo, cuando diez documentos son considerados, los resultados son lo suficientemente buenos como para asegurar que la estrategia empleada para la generación de resúmenes está siendo efectiva.

Los resultados que logran ser notorios en la figura 5.1 son los obtenidos por medio de los MRFs. Pues al contrario de el documento virtual y el resumen multi-documento, los MRF logran desempeñarse de mejor manera cuando se dan como retroalimentación pocos *documentos ejemplo*. De manera particular, la configuración de un MRF con 2

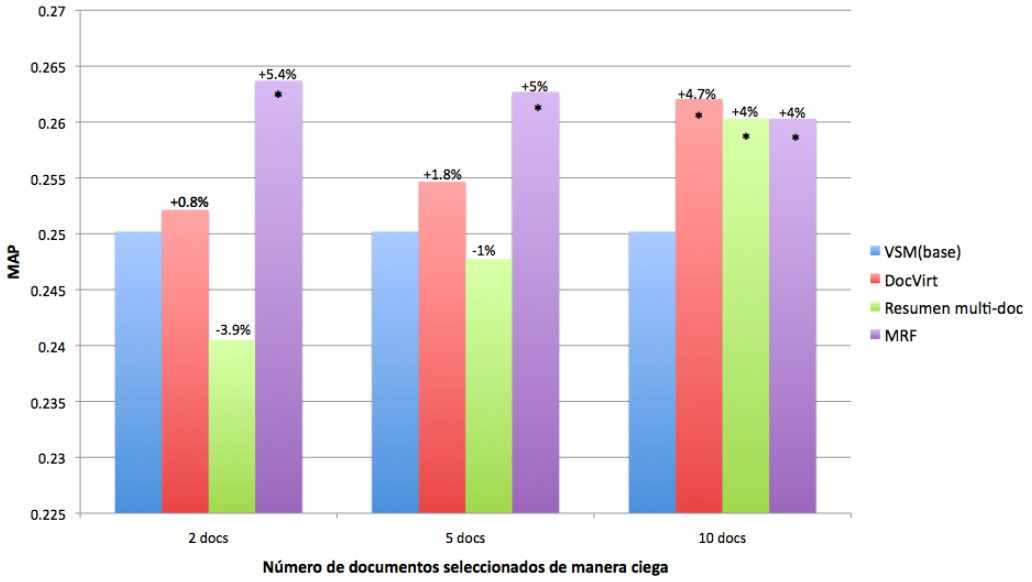


**Figura 5.1: Mejores resultados obtenidos empleando un proceso de selección ciega**

documentos ejemplo es la que logra obtener los mejores resultados bajo el esquema de selección ciega, logrando superar al método base en un 5.4 %.

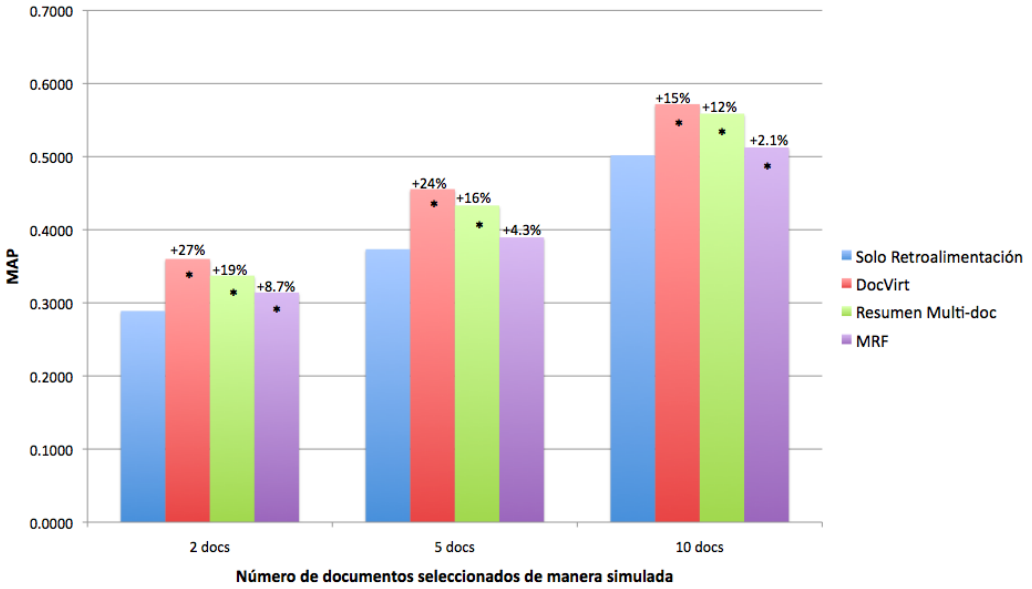


**Figura 5.2: Mejores resultados obtenidos empleando un proceso de selección simulada**

La figura 5.2 muestra de manera resumida los resultados obtenidos por las diferen-

tes estrategias de re-ordenamiento propuestas, considerando un esquema de selección simulada para los *documentos ejemplo*. De igual manera se marcan los resultados que logran ser significativos con un (*) así como se detalla el porcentaje de ganancia o pérdida de cada configuración. En esta figura (5.2) los resultados mostrados para el documento virtual y el resumen multi-documento corresponden a una forma de representación *temática* y *geográfica simple*, mientras que para el caso de los MRFs la configuración que se muestra es la correspondiente a la *geográfica simple*. Es conveniente recordar que estos experimentos son comparados contra el proceso único de la retroalimentación, es decir, las columnas marcadas con la etiqueta "Sólo Retroalimentación" representan el comportamiento del sistema tras colocar al principio de la lista de documentos recuperados a los $n$ elementos seleccionados por el proceso de retroalimentación simulada.

Para los resultados obtenidos manejando un documento virtual y un resumen multi-documento es posible observar que el orden generado es bastante mejor que el obtenido por el proceso de retroalimentación. Para el caso de un documento virtual basta con dos documentos ejemplo para lograr una mejora del 27 %. De manera similar, cuando se maneja un resumen multi-documento bastan 2 documentos para lograr una mejora del 19 % con respecto al proceso de sólo retroalimentación. Aunque existe una diferencia notable entre los resultados obtenidos con el documento virtual y el resumen multi-documento, ambos generan resultados significativos, por lo cual podemos decir que la perdida en precisión del resumen multi-documento es aceptable dado que sólo contienen un 40 % de la información de los documentos ejemplo.

Para el caso de los resultados obtenidos con los MRF's es notorio que la ganancia es mínima comparada contra el proceso de sólo la retroalimentación. La razones de este comportamiento tienen mucho que ver con la naturaleza del algoritmo de optimización empleado, ICM, el cual es un algoritmo rápido pero con la desventaja de que puede caer máximos locales. Ahora bien, si tomamos en cuenta que los conjuntos de documentos son muy parecidos entre si, es decir, elementos relevantes y no relevantes son muy similares, y dado que la función de energía toma en cuenta este tipo de información para determinar si un elemento es relevante o no relevante, esto no ayuda en el proceso de encontrar la configuración óptima del MRF.

Con respecto a las formas de representar a los documentos, en general las representaciones puramente geográficas fueron las mas deficientes. De manera particular se observó que la forma más deficiente para tal propósito es la que emplea *coordenadas*

*geográficas*. Para tratar de determinar el por qué sucedía esto se hizo un análisis a varios conjuntos de documentos considerados relevantes a consultas geográficas. Tras esto se observó que los focos geográficos de estos documentos se encuentran muy dispersos en el mapa (i.e., coordenadas geográficas muy distantes); razón por la que este tipo de represetación es la más deficiente. Por ejemplo, la consulta "Attacks in Japanesse subways" tiene un total de 54 documentos relevantes, dentro de los cuales se hace mención a gran cantidad de lugares que no están cerca de Japón, que es el foco principal de la consulta. Esto queda expresado de mejor manera en la figura 5.3, donde los puntos dibujados en el mapa muestran los lugares mencionados en el conjunto de los 54 documentos relevantes de la consulta.


**Figura 5.3: Puntos geográficos mencionados dentro de los documentos relevantes a la consulta GC-93:Attacks in Japanese subways**

Debido a este mismo fenómeno es que la representación *geográfica expandida* tampoco aporta muchas ventajas. Pues aunque los procesos de desambiguación y de expansión sean lo suficientemente precisos, en los casos como el mostrado en la figura 5.3 los términos agregados no serán lo suficientemente discriminativos para el proceso del re-ordenamiento y por el contrario pueden estar agregando términos ruidosos.

En general, se observó que una forma *combinada*, i.e., temática y geográfica, es la mejor forma de representar a los documentos, pues en estos casos es cuando mejores

resultados se obtienen. En este punto es importante recalcar que en general cuando $\lambda$ es 0.6 es cuando mejor desempeño se puede obtener, lo cual quiere decir, que los elementos que ayudan mejor a la discriminación entre elementos relevantes y no relevantes siguen siendo los elementos temáticos de los documentos y no los geográficos.

Con la finalidad de conocer el por qué los elementos temáticos son los elementos más discriminativos se hizo un análisis sobre las consultas del GeoCLEF así como de la colección de datos. De este análisis fue posible observar que en la mayoría de las consultas (60 % aprox) los requerimientos de información temática hacen mención a eventos o actos muy concretos; por ejemplo las consultas listadas a continuación:

```
GC-14 : Environmentally hazardous Incidents in the North Sea
GC-55 : Deaths caused by avalanches occurring in Europe, but
not in the Alps
GC-99 : Floods in European cities
```

Observe que la restricción geográfica en estas consultas queda muy abierta, razón por la que los elementos temáticos son los más discriminativos al momento de estar realizando el ordenamiento. De manera particular para consultas de este tipo fue cuando el dar mayor peso a la parte temática permite obtener un mejor desempeño al método de re-ordenamiento.

En un porcentaje menor se tienen consultas donde la parte geográfica juega realmente un papel importante. Ejemplo de estas son:

```
GC-22 : Restored buildings in Southern Scotland
GC-64 : Sport events in the french speaking part of
Switzerland
```

Nótese que en estas consultas la parte temática es menos precisa que la parte geográfica. Para consultas de este tipo juega un papel importante la tarea de desambiguación geográfica, pues dado que la restricción geográfica es muy específica, el resolver adecuadamente su ambigüedad permitirá resolver en mejor medida las necesidades de información del usuario. En los experimentos realizados fue posible notar que en las consultas que tienen estas características es conveniente dar mayor peso a la parte geográfica de los documentos en el proceso de re-ordenamiento.

En conclusión podemos decir que si las consultas tienen requerimientos de información temática muy específicos y una restricción geográfica muy general, es conveniente dar mayor importancia ($\lambda$ grande) a la parte temática de los documentos para poder generar un orden más pertinente. Por otro lado, si la restricción geográfica es muy específica y las necesidades de información temática muy generales, es conveniente dar mayor importancia ($\lambda$ grande) a la parte geográfica de los documentos en el proceso de re-ordenamiento.

# Capítulo 6
# Conclusiones

En este trabajo de investigación se presentó una nueva estrategia de re-ordenamiento o *re-ranking*, como se conoce en Inglés, para el problema de Recuperación de Información Geográfica. La idea de enfocarnos en el problema del ordenamiento y no en el problema de la recuperación surge debido al exhaustivo estudio de los diferentes sistemas hasta ahora propuestos por diferentes grupos de investigación, de donde se concluyó que el problema GIR es parcialmente resuelto por máquinas robustas de IR, pues logran obtener niveles aceptables de recuerdo (al rededor del 80 %), sin embargo presentan problemas al momento de generar un orden pertinente de éstos antes de ser entregados al usuario, es decir tienen baja precisión.

Las deficiencias de las máquinas GIR o incluso de máquinas tradicionales de IR para generar un orden pertinente de la lista de documentos recuperados tiene su origen en la falta de información. Es decir, la información de la consulta resulta ser insuficiente para que la máquina de recuperación pueda otorgar un orden adecuado, y si a esto le agregamos el hecho de que las consultas geográficas tienden a contener requisitos de información en forma implícita, el problema se vuelve aún más difícil de resolver. De estas observaciones es de donde surge la idea principal de este trabajo, que involucró el utilizar *ejemplos* de lo que el usuario está buscando en lugar de sólo términos aislados para tratar de responder a las necesidades de información de éste. En el fondo, los *documentos ejemplo* tienen como propósito el proporcionar al módulo de re-ordenamiento una aproximación en forma más explícita de los requerimientos de información del usuario.

De esta forma, nuestra arquitectura propuesta considera como módulo principal el proceso de re-ordenamiento, el cual recibe una lista de documentos recuperados (proveniente de cualquier máquina de IR) junto con un conjunto de *documentos ejemplo* los cuales son obtenidos a través de un proceso de selección (retroalimentación). La

información obtenida a través de este proceso se emplea para hacer el re-ordenamiento de los documentos recuperados, y no para hacer un segundo proceso de recuperación, aportando rapidez y evitando las desventajas que tienen los procesos de expansión de consultas vía retroalimentación de relevancia [93]. Así, la aportación principal de este trabajo es un sistema de re-ordenamiento vía retroalimentación de relevancia en el contexto de búsquedas geográficas.

El método propuesto mostró ser mejor al momento de generar un orden de los documentos, superando al método base en un 5.4 % en el mejor de los casos bajo un esquema de selección ciega, mientras que para el caso de selección simulada se logró tener una ventaja de hasta un 27 % con respecto al método base. Dado que para atacar el problema del re-ordenamiento se propusieron diferentes estrategias, a continuación se da un breve recordatorio de cada estrategia junto con la principal conclusión a la que se llegó.

El primer problema al que nos enfrentamos fue al de cómo manejar la información contenida en los *documentos ejemplo*, para lo cual tres ideas fueron propuestas y evaluadas:

- Documento Virtual. Esta idea propone la construcción de un documento virtual a partir de toda la información contenida en los *documentos ejemplo* con la finalidad de medir la similitud de cada uno de los elementos de la lista original de documentos contra este documento virtual y de esta forma proporcionar un orden más pertinente para lista de documentos. De los experimentos realizados se concluyó que la propuesta de re-ordenamiento empleando un documento virtual permite en efecto proporcionar un ordenamiento más pertinente. En otras palabras, el emplear los ejemplos para el proceso del re-ordenamiento permite interpretar de mejor manera los requerimientos de información del usuario. Para este caso, la mejor configuración logra superar al método base en un 4.7 % bajo un esquema de selección ciega, mientras que para una selección simulada el mejor resultado mejora en un 27 % al método base.

- Resumen Multi-Documento. Esta idea propone la construcción de un documento corto (resumen) a partir del conjunto de *documentos ejemplo*. Dado que se supone son ejemplos relacionados temáticamente, se espera que el resumen contenga sólo la información relevante de éstos, así como ayudar a resolver redundancias entre ellos. Los mejores resultados mejoran al método base en un

4 % y un 19 % considerando un esquema de selección ciega y simulada respectivamente. Dado que estos resultados mostraron ser significativos, se concluye que la perdida de precisión en comparación con los resultados obtenidos empleando un documento virtual es aceptable dado que el tamaño del resumen es de sólo el 40 % del tamaño de los *documentos ejemplo*.

- Campo Aleatorio de Markov. La idea del campo de Markov es integrar tanto la información proporcionada por la máquina de IR, i.e. el orden originalmente dado a los documentos, la similitud entre los documentos recuperados y la intención de búsqueda (dada por los *documentos ejemplo*) con la finalidad de generar un orden más apropiado para la lista de documentos recuperados. Los experimentos realizados mostraron que bajo un esquema de selección ciega considerando pocos documentos, el MRF es capaz de otorgar un mejor ordenamiento de los documentos, mejorando en un 5.7 % al método base. Para el caso de un esquema de selección simulada, los resultados apenas logran mejorar mejorar en un 8.7 % al método base. Las conclusiones principales de estos experimentos son: *i)* los MRF's se comportan mejor con pocos documentos, i.e., manejando poca información, lo cual permite a la función de energía diferenciar mejor a elementos relevantes de aquellos que no lo son, *ii)* formas de representación de los documentos que permitan aumentar las distancias entre ellos permite al MRF aproximarse a la configuración óptima, en particular una forma de representación puramente geográfica.

El segundo problema al que nos enfrentamos después de definir la forma en cómo se manejaría la información de los *documentos ejemplo* fue al de cómo representar esta información de tal forma que fuera posible medir la similitudes entre documentos. Para esto se propusieron siete formas distintas de representar a los documentos, la primera y más intuitiva fue la de emplear a las palabras contenidas en los documentos, forma de representación tipo *BOW*. De manera adicional, dado que el contexto del trabajo son búsquedas geográficas, tres formas de representación geográficas fueron propuestas: *i)* entidades geográficas simples, *ii)* entidades geográficas expandidas, y *iii)* coordenadas geográficas. Finalmente una forma de representación combinada (temática y geográfica) fue propuesta con la finalidad de evaluar el grado de complementariedad que tienen ambas partes de información. En esta forma de representación combinada, la parte geográfica se representó de las tres formas mencionadas anterior-

mente.

De los experimentos realizados manejando estas diferentes formas de representación se concluyó que la forma más apropiada para representar a los documentos es por medio de una forma combinada. Se mostró que las partes temática y geográfica de los documentos son complementarias, y por sí solas no son capaces de representar adecuadamente a los documentos, principalmente las que aportan menores beneficios son las formas de representación puramente geográficas. De manera particular, empleando una forma de representación combinada y con un valor de $\lambda$ =0.6 es cuando el mejor compromiso se alcanza entre la parte temática y la parte geográfica de los documentos. La razón de este valor es debido a la naturaleza de las consultas, pues en su mayoría contienen requerimientos de información temática muy específica.

Un punto interesante que se observó durante los experimentos fue cuando los documentos se representaron por medio de coordenadas geográficas, pues aunque era de suponerse que los documentos relevantes estarían cercanos geográficamente, en la realidad esto no estaba sucediendo, siendo esta la principal razón por la que las formas de representación geográficas, y en particular la de coordenadas geográficas no permitieran obtener resultados favorables. Esto se terminó de confirmar, cuando al revisar los resultados obtenidos combinando fuentes de información (i.e., partes temática y geográfica) se notó que los mejores resultados tendían a obtenerse cuando mayor importancia se le daba a la parte temática de los documentos. De aquí otra de nuestras conclusiones principales que dice que la colección de documentos así como las consultas provistas por el foro del GeoCLEF no son completamente geográficas, pues elementos altamente discriminativos siguen siendo elementos temáticos.

## 6.1. Trabajo Futuro

Tomando en cuenta las restricciones y/o desventajas que presenta el método propuesto, algunas ideas que se pretende explorar en un futuro son:

- Explorar técnicas automáticas para la selección de los *documentos ejemplo* que permitan obtener un buen desempeño del sistema. La idea que se tiene en este punto es proponer un conjunto de características para representar a un sub-conjunto del total de documentos recuperados, de tal forma que se pueda determinar de una manera automática qué documentos podrían ser buenos candidatos para ser *documentos ejemplo*.

- Trabajar en la construcción y configuración de un método de generación de resúmenes orientado a lo geográfico y no a lo temático. De esta forma se esperaría que tras hacer el resumen el foco geográfico de los *documentos ejemplo* quede mejor acotado y de esta forma proporcionar un re-ordenamiento más preciso.

- Emplear un algoritmo de optimización diferente al ICM en los MRFs, de tal forma que sea posible lograr una configuración óptima del campo.

- El no contar con una máquina de IR especializada en el contexto de búsquedas geográficas nos hace depender en gran medida de la calidad de la salida otorgada por la máquina de IR empleada en el proceso de recuperación, pues como se dijo antes, nuestro trabajo parte de la suposición de que suficientes elementos relevantes han sido recuperados. Como trabajo adicional se propone el hacer experimentos con diferentes máquinas de IR con la finalidad de evaluar la generalidad del método propuesto al determinar el grado de dependencia o independencia de la máquina de IR.

- Definir una forma de medir el grado de especificidad de los elementos temáticos y geográficos de las consultas con la finalidad de escoger en el proceso del re-ordenamiento la mejor configuración posible.

- Finalmente, dado que los resultados mostraron que el método propuesto se desempeña adecuadamente con elementos temáticos, se espera poder hacer pruebas de la estrategia de re-ordenamiento en colecciones de consultas y de documentos diferentes, con la finalidad de determinar la estabilidad y generalidad del método propuesto más allá del contexto GIR.

## 6.2. Publicaciones derivadas de la investigación

**Publicaciones en Revistas Arbitradas**

- A Probabilistic Method for Ranking Refinement in Geographic Information Retrieval. Esaú Villatoro-Tello, R. Omar Chavéz-García, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda, L. Enrique Sucar. Procesamiento de Lenguaje Natural, No. 44, pp. 123-130, April 2010.

**Publicaciones en Conferencias Arbitradas**

- Ranking Refinement via Relevance Feedback in Geographic Information Retrieval. Esaú Villatoro-Tello, Luis Villaseñor-Pineda, and Manuel Montes-y-Gómez. Mexican International Conference on Artificial Intelligence MICAI 2009. Guanajuato, Mexico, November 2009. Lecture Notes in Artificial Intelligence Vol. 5845, Springer, 2009.

- Multi-Document Summarization Based on Locally Relevant Sentences. Esaú Villatoro-Tello, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez, and David Pinto-Avendaño. Mexican International Conference on Artificial Intelligence MICAI 2009. Guanajuato, Mexico, November 2009. IEEE, 2009.

- A Ranking Approach based on Sample Documents for Geographic Information Retrieval. Esaú Villatoro-Tello, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda. Post-proccedings of the 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008. Lecture Notes in Computer Science Vol. 5706, Springer 2009.

**Publicaciones en Conferencias no Arbitradas**

- INAOE at GeoCLEF 2008: A Ranking Approach based on Sample Documents. Esaú Villatoro-Tello, Manuel Montes-y-Gómez, and Luis Villaseñor-Pineda. Working Notes of CLEF 2008. Aarhus, Denmark. September 2008.

**Otras Publicaciones**

- Recuperación de Información aplicando un Ordenamiento Basado en Documentos Ejemplo. Esaú Villatoro-Tello, Luis Villaseñor-Pineda, y Manuel Montes-y-Gómez. Memorias del 9o Encuentro de Investigación del INAOE. Puebla, México. Noviembre 2008.

# Bibliografía


[1] Cross-lingual evaluation forum. http://www.clef-campaign.org/, September 2010.

[2] Text retrieval conference (trec). http://trec.nist.gov/, September 2010.

[3] M. Agosti, F. Crivellari, G. Deambrosis, and G. Gradenigo. An architecture and design approach for a geographic information retrieval system to support retrieval by content and browsing. *Computer, Environment and urban Systems*, 17(4):321–335, 1993.

[4] Geoffrey Andogah and Gosse Bouma. Relevance measures using geographic scopes and types. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science*, pages 794–801. Springer, 2008.

[5] Leonardo Andrade and Mário J. Silva. Relevance ranking for geographic ir. In *Proceedings of 3rd Workshop on Geographic Information Retrieval, GIR-2006, SIGIR*, Seattle, WA, USA, 2006. ACM Press. Extended abstract.

[6] Avi Arampatzis, Marc van Kreveld, Iris Reinbacher, Christopher B. Jones, Subodh Vaid, Paul Clough, Hideo Joho, Marc Benkert, and Alexander Wolff. Web-based delineation of imprecise regions. In *Proceedings of Workshop on Geographic Information Retrieval, SIGIR 04*, Sheffield, UK, 2004. Extended abstract.

[7] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrival*. Addison Wesley, 1999.

[8] Karla A. Borges, Alberto H. F. Laender, Claudia B. Medeiros, and Clodoveu A. Davis Jr. Discovering geographic locations in web pages using urban addresses. In *Proceedings of 4th InternationalWorkshop on Geographic Information Retrieval GIR 2007, ACM CIKM-2007*, pages 31–36, Lisbon, Portugal, 2007. ACM Press.

[9] Davide Buscaldi, José M. Perea-Ortega, Paolo Rosso, Luis Alfonso Ureña López, Daniel Ferrés, and Horacio Rodríguez. Geotextmess: Result fusion with fuzzy borda ranking in geographical information retrieval. In *Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008*, volume 5706 of *Lecture Notes in Computer Science*, pages 867–874. Springer, 2009.

[10] Davide Buscaldi and Paolo Rosso. A comparision of methods for the automatic identification of locations in wikipedia. In *Proceedings of 4th International Workshop on Geographic Information Retrieval GIR 2007, ACM CIKM-2007*, pages 89–91, 2007.

[11] Davide Buscaldi and Paolo Rosso. Geo-wordnet: Automatic georeferencing of wornet. In *Proc. 6th Int. Conf. on Language Resources and Evaluation, LREC-2008*, Marrakech, Morocco, 2008.

[12] Davide Buscaldi and Paolo Rosso. Map-based vs. knowledge-based toponym disambiguation. In *Proceedings of 5th International Workshop on Geographic Information Retrieval, GIR 2008, CIKM-2008*, pages 19–22, Napa Valley, U.S.A., 2008.

[13] Davide Buscaldi and Paolo Rosso. On the relative importance of toponyms in geoclef. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science*, pages 815–822. Springer, 2008.

[14] Guoray Cai. Contextualization of geospatial database semantics for human-gis interaction. *Geo-Informatica*, 11(2):217–237, June, 2007.

[15] Nuno Cardoso, Davis Cruz, Marcirio Chaves, and Mário J. Silva. Using geographic signatures as query and document scopes in geographic ir. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the*

*Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science*, pages 802–810. Springer, 2008.

[16] Nuno Cardoso, Bruno Martins, Marcirio Silveira Chaves, Leonardo Andrade, and Mário J. Silva. The xldb group at geoclef 2005. In *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*, volume 4022 of *Lecture Notes in Computer Science*, pages 997–1006. Springer, 2006.

[17] Nuno Cardoso, Patricia Sousa, and Mário J. Silva. Experiments with geographic evidence extracted from documents. In *Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008*, volume 5706 of *Lecture Notes in Computer Science*, pages 885–893. Springer, 2009.

[18] Rama Chellappa. *Markov Random Fields: Theory and Application*. Boston: Academic Press, 1993.

[19] Omar Chávez, Enrique Sucar, and Manuel Montes. Image re-ranking based on relevance feedback combining internal and external similarities. In *The FLAIRS Conference*, Daytona Beach, Florida, USA, 2010.

[20] Paul Clough and Mark Sanderson. A proposal for comparative evaluation of automatic annotation for geo-referenced documents. In *Proceedings of Workshop on Geographic Information Retrieval, SIGIR 04*, Sheffield, UK, 2004. Extended abstract.

[21] Paolo Rosso Davide Buscaldi and Piedachu Peris Garcia. Inferring geographical ontologies from multiple resources for geographical information retrieval. In *Proceedings of 3rd Workshop on Geographic Information Retrieval, GIR-2006, SIGIR*, Seattle, WA, USA, 2006. ACM Press. Extended abstract.

[22] Junyan Ding, Luis Gravano, and Narayanan Shivakumar. Computing geographical scopes of web resources. In *Proceedings of the 26th International Conference on Very Large Data Bases, VLDB'00*, pages 545–556, San Francisco, CA, USA, 2000.

[23] Óscar Ferrández, Zornitsa Kozareva, Antonio Toral, Elisa Noguera, Andrés Montoyo, Rafael Muñoz, and Fernando Llopis. University of alicante at geoclef 2005. In *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*, volume 4022 of *Lecture Notes in Computer Science*, pages 924–927. Springer, 2006.

[24] Daniel Ferrés, Alicia Ageno, and Horacio Rodríguez. The geotalp-ir system at geoclef 2005: Experiments using a qa-based ir system, linguistic analysis, and a geographical thesaurus. In *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*, volume 4022 of *Lecture Notes in Computer Science*, pages 947–955. Springer, 2006.

[25] Daniel Ferrés and Horacio Rodríguez. Talp at geoclef 2007: Results of a geographical knowledge filtering approach with terrier. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science*, pages 830–833. Springer, 2008.

[26] Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, and Peter Yanker. Query by image and video content: The qbic system. *IEEE Computer Special Issue on Content-Based Retrieval*, 28(9):23–32, September 1995.

[27] Gaihua Fu, Christopher B. Jones, and Alia I. Abdelmoty. Ontology-based spatial query expansion in information retrieval. In *On the move to meaningful Internet Systems 2005: COOPIS, DOA, and ODBASE*, volume 3761 of *Lecture Notes in Computer Science*, pages 1466–1482. Springer, 2005.

[28] Salton G. and Buckley C. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[29] Salton G., Yang C. S., and A. Wong. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[30] Manuel García-Vega, Miguel A. García-Cumbreras, L. Alfonso Ureña-López, and José M. Perea-Ortega. Geouja system. the first participation of the university of

jaén at geoclef 2006. In *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, volume 4730 of *Lecture Notes in Computer Science*, pages 913–917. Springer, 2007.

[31] Fredric Gey, Ray Larson, Mark Sanderson, Hideo Joho, Paul Clough, and Vivien Petras. Geoclef: the clef 2005 cross language geographic information retrieval track overview. In *Working notes for the CLEF 2005 Workshop*, Wien, Austria, September 2005.

[32] Fredric C. Gey, Ray R. Larson, Mark Sanderson, Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Giorgio Maria Di Nunzio, and Nicola Ferro. Geoclef 2006: The clef 2006 cross-language geographic information retrieval track overview. In *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, volume 4730 of *Lecture Notes in Computer Science*, pages 852–876. Springer, 2007.

[33] Rocio Gillén. Monolingual and bilingual experiments in geoclef2006. In *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, volume 4730 of *Lecture Notes in Computer Science*, pages 893–900. Springer, 2007.

[34] Miguel Á. Gracía-Cumbreras, José M. Perea-Ortega, Manuel García-Vega, and L. Alfonso Ureña-López. Information retrieval with geographical references. relevant documents filtering vs. query expansion. *Information Processing and Management*, 45(5):605–614, 2009.

[35] Jens Graupmann and Ralf Schenkel. Geospheresearch: Context-aware geographic web search. In *Proceedings of 3rd Workshop on Geographic Information Retrieval, GIR-2006, SIGIR*, Seattle, WA, USA, 2006. ACM Press. Extended abstract.

[36] David A. Grossman and Ophir Frieder. *Information Retrieval, Algorithms and Heuristics*. Springer, second edition, 2004.

[37] Rocio Guillén. Csusm experiments in geoclef 2005: Monolingual and bilingual tasks. In *Accessing Multilingual Information Repositories: 6th Workshop of the*

*Cross-Language Evaluation Forum, CLEF 2005*, volume 4022 of *Lecture Notes in Computer Science*, pages 956–962. Springer, 2006.

[38] Rocio Guillén. Geoparsing web queries. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science*, pages 781–785. Springer, 2008.

[39] Claudia Hauff, Dolf Trieschnigg, and Henning Rode. University of twente at geoclef 2006: geofiltered document retrieval. In *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, volume 4730 of *Lecture Notes in Computer Science*, pages 958–961. Springer, 2007.

[40] Andreas Henrich and Volker Lüdecke. Characteristics of geographic information needs. In *Proceedings of 4th Internationa lWorkshop on Geographic Information Retrieval GIR 2007, ACM CIKM-2007*, pages 1–6, Lisbon, Portugal, 2007. ACM Press.

[41] Linda L. Hill, Michael F. Goodchild, and Greg Jane. Research directions in geographic ir based on the alexandria digital library project. In *Proceedings of Workshop on Geographic Information Retrieval, SIGIR'04*, Sheffield, UK, 2004. Extended abstract.

[42] Eduard Hovy. *The Oxford Handbook of Computational Linguistics*, chapter Text Summarization, pages 582–598. Oxford, 2003.

[43] Eduard Hovy and Chin-Yen Lin. *Advances in Automatic Text Summarization*, chapter Automated text summarization in SUMMARIST, pages 81–94. MIT Press, Cambridge, 1999.

[44] You-Heng Hu and Linlin Ge. The university of new south wales at geoclef 2006. In *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, volume 4730 of *Lecture Notes in Computer Science*, pages 905–912. Springer, 2007.

[45] You-Heng Hu and Linlin Ge. Learning ranking functions for geographic information retrieval using genetic programming. *Journal of Research and Practice in Information Technology*, 41(1):39–52, 2009.

[46] B. Jones, Harith Alani, and Douglas Tudhope. Geographical information retrieval with ontologies of place. *International Journal of Geographical Information Science*, 15(4):287–306, 2001.

[47] Christopher B. Jones, Harith Alani, and Douglas Tudhope. Geographical information retrieval with ontologies of place. In *Working notes for the CLEF 2006 Workshop*, Alicante, Spain, September 2006.

[48] Christopher B. Jones and Ross S. Prurves. Geographical information retrieval. *International Journal of Geographic Information Science*, 22(3):219–228, 2008.

[49] R. Kindermann. Markov random field and their applications. *Contemporary Mathematics, V:1*, 1980.

[50] Sara Lana-Serrano, José M. Goñi-Menoyo, and José C. González-Cristóbal. Miracle at geoclef 2005: First experiments in geographical ir. In *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*, volume 4022 of *Lecture Notes in Computer Science*, pages 920–923. Springer, 2006.

[51] Sara Lana-Serrano, José M. Goñi-Menoyo, and José C. González-Cristóbal. Report of miracle team for geographic ir in clef 2006. In *Working notes for the CLEF 2006 Workshop*, Alicante, Spain, September 2006.

[52] Ray R. Larson. Cheshire ii at geoclef 2005: Fusion and query expansin for gir. In *Working notes for the CLEF 2005 Workshop*, Wien, Austria, September 2005.

[53] Ray R. Larson. Cheshire at geoclef 2007: Retesting text retrieval baselines. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science*, pages 811–814. Springer, 2008.

[54] Ray R. Larson. Cheshire at geoclef 2008: Text and fusion approaches for gir. In *Evaluating Systems for Multilingual and Multimodal Information Access, 9th*

*Workshop of the Cross-Language Evaluation Forum, CLEF 2008*, volume 5706 of *Lecture Notes in Computer Science*, pages 830–837. Springer, 2009.

[55] Ray R. Larson and Patricia Frontiera. Ranking and representation for geographical information retrieval. In *Proceedings of Workshop on Geographic Information Retrieval SIGIR 04*, Sheffield, UK, 2004. Extended abstract.

[56] Ray R. Larson, Fredric C. Gey, and Vivien Petras. Berkeley at geoclef: Logistic regression and fusion for geographic information retrieval. In *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*, volume 4022 of *Lecture Notes in Computer Science*, pages 963–976. Springer, 2006.

[57] Joon Ho Lee. Analysis of multiple evidence combination. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–276, 1997.

[58] Jochen L. Leidner. Experiments with geo-filtering predicates for ir. In *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*, volume 4022 of *Lecture Notes in Computer Science*, pages 987–996. Springer, 2006.

[59] Johannes Leveling and Sven Hartrumpf. On metonymy recognition for geographic ir. In *Proceedings of 3rd Workshop on Geographic Information Retrieval, GIR-2006, SIGIR*, Seattle, WA, USA, 2006. ACM Press. Extended abstract.

[60] S. Li. Markov random field models in computer vision. *Computer Vision at ECCV '94*, pages 361–370, 1994.

[61] Yi Li, Alistair Moffat, Nicola Stokes, and Lawrence Cavedon. Exploring probabilistic toponym resolution for geographic information retrieval. In *Proceedings of 3rd Workshop on Geographic Information Retrieval, GIR-2006, SIGIR*, Seattle, WA, USA, 2006. ACM Press. Extended abstract.

[62] Zhisheng Li, Chong Wang, Xing Xie, and Wei-Ying Ma. Msra columbus at geoclef 2006. In *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, volume 4730 of *Lecture Notes in Computer Science*, pages 926–929. Springer, 2007.

[63] Zhisheng Li, Chong Wang, Xing Xie, and Wei-Ying Ma. Exploring lda-based document model for geographic information retrieval. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science*, pages 842–849. Springer, 2008.

[64] Wen-Cheng Lin and Hsin-Hsi Chen. Merging mechanisms in multilingual information retrieval. In *Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002*, volume 2785 of *Lecture Notes in Computer Science*, pages 175–186. Springer, 2003.

[65] Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker, and Xing Xie. Geoclef 2007: The clef 2007 cross language geographic information retrieval track overview. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science*, pages 745–772. Springer, 2008.

[66] Bruno Martins, Nuno Cardoso, Marcirio Silveira Chaves, Leonardo Andrade, and Mário J. Silva. The university of lisbon at geoclef 2006. In *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, volume 4730 of *Lecture Notes in Computer Science*, pages 986–994. Springer, 2007.

[67] Bruno Martins, Mário J. Silva, and Leonardo Andrade. Indexing and ranking in geo-ir systems. In *Proceedings of Workshop on Geographic Information Retrieval, GIR'05*, pages 31–34, Bremen, Germany, 2005. ACM Press.

[68] Bruno Martins, Mário J. Silva, Sérgio Freitas, and Ana P. Afonso. Handling locations in search engine queries. In *Proceedings of the 3rd Workshop on Geographic Information Retrieval, GIR'06*, Seattle, WA, USA, 2006. ACM Press. Extended abstract.

[69] Felix Mata. irank: Ranking geographical information by conceptual, geographic and topologic similarity. In *GeoS '09: Proceedings of the 3rd International Conference on GeoSpatial Semantics*, pages 159–174. Springer-Verlag, 2009.

[70] Simon E. Overell, João Magalhães, and Stefan Rüger. Gir experiments with forostar. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science*, pages 856–863. Springer, 2008.

[71] Simon E. Overell, Adam Rae, and Stefan Rüger. Geographic and textual data fusion in forostar. In *Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008*, volume 5706 of *Lecture Notes in Computer Science*, pages 838–842. Springer, 2009.

[72] Simon E. Overell and Stefan Rüger. Identifying and grounding descriptions of places. In *Proceedings of 3rd Workshop on Geographic Information Retrieval, GIR'06*, Seattle, WA, USA, 2006. ACM Press. Extended abstract.

[73] Simon E. Overell and Stefan Rüger. Geographic co-occurrence as a tool for gir. In *Proceedings of 4th International Workshop on Geographic Information Retrieval GIR 2007, ACM CIKM-2007*, pages 71–76, Lisbon, Portugal, 2007. ACM Press.

[74] José M. Perea-Ortega, Miguel A. García Cumbreras, Manuel García-Vega, and Arturo Montejo-Ráez. Geouja system. university of jaén at geoclef 2007. In *Working notes for the CLEF 2007 Workshop*, Budapest, Hungary, September 2007.

[75] José M. Perea-Ortega, Miguel A. García-Cumbreras, Miguel Garca-Vega, and L. Alfonso Ureña-López. Sinai-gir system. university of jaén at geoclef 2008. In *Working notes for the CLEF 2008 Workshop*, Aarhus, Denmark, September 2008.

[76] M. F. Porter. An algorithm for suffix stripping. pages 313–316. Morgan Kaufmann Publishers Inc., 1997.

[77] Mark Sanderson and Yu Han. Search words and geography. In *Proceedings of 4th International Workshop on Geographic Information Retrieval GIR 2007, ACM CIKM-2007*, pages 13–14, Lisbon, Portugal, 2007. ACM Press.

[78] Mark Sanderson and Janet Kohler. Analyzing geographic queries. In *Proceedings of Workshop on Geographic Information Retrieval SIGIR 04*, Sheffield, UK, 2004. Extended abstract.

[79] Steven Schockaert, Martine De Cock, and Etienne E. Kerre. Towards fuzzy spatial reasoning in geographic ir systems. In *Proceedings of 3rd Workshop on Geographic Information Retrieval, GIR-2006, SIGIR*, Seattle, WA, USA, 2006. ACM Press. Extended abstract.

[80] Mário J. Silva, Bruno Martins, Marcirio Chaves, Nuno Cardoso, and Ana Paula Afonso. Adding geographic scopes to web resources. In *Proceedings of Workshop on Geographic Information Retrieval, SIGIR'04*, Sheffield, UK, 2004. Extended abstract.

[81] Alberto H. F. Laender Tiago M. Delboni, Karla A. V. Borges. Geographic web search based on positioning expressions. In *Proceedings of Workshop on Geographic Information Retrieval, GIR'05*, pages 61–64, Bremen, Germany, 2005.

[82] Antonio Toral, Óscar Ferrández, Elisa Noguera, Zornitsa Kozareva, Andrés Montoyo, and Rafael Muñoz. Gir with geographic query expansion. In *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, volume 4730 of *Lecture Notes in Computer Science*, pages 889–892. Springer, 2007.

[83] Manuel García Vega, Miguel Angel García Cumbreras, Luis Alfonso Ureña López, José M. Perea-Ortega, Francisco Javier Ariza-López, Óscar Ferrández, Antonio Toral, Zornitsa Kozareva, Elisa Noguera, Andrés Montoyo, Rafael Muñoz, Davide Buscaldi, and Paolo Rosso. R2d2 at geoclef 2006: A combined approach. In *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, volume 4730 of *Lecture Notes in Computer Science*, pages 918–925. Springer, 2007.

[84] Esaú Villatoro-Tello, R. Omar Chavéz-García, Manuel Montes y Gómez, Luis Villaseñor-Pineda, and L. Enrique Sucar. A probabilistic method for ranking refinement in geographic information retrieval. *Procesamiento de Lenguaje Natural*, 44:123–130, April 2010.

[85] Esaú Villatoro-Tello, Luis Villaseñor-Pineda, and Manuel Montes y Gómez. Using word sequences for text sumarization. In *Proceedings of the 9th International Conference on Text, Speech and Dialogue (TSD 2006)*, pages 293–300, Brno, Czech Republic, 2006.

[86] Esaú Villatoro-Tello, Luis Villaseñor-Pineda, Manuel Montes y Gómez, and David Pinto-Avendaño. Multi-document summarization based on locally relevant sentences. pages 87–91. IEEE Computer Society, 2009.

[87] Esaú Villatoro-Tello, Manuel Montes y Gómez, and Luis Villaseñor-Pineda. A ranking approach based on example texts for geographic information retrieval. In *Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008*, volume 5706 of *Lecture Notes in Computer Science*, pages 875–879. Springer, 2009.

[88] Esaú Villatoro-Tello, Manuel Montes y Gómez, and Luis Villaseñor-Pineda. Ranking refinement via relevance feedback in geographic information retrieval. In *MICAI '09: Proceedings of the 8th Mexican International Conference on Artificial Intelligence*, pages 165–176. Springer-Verlag, 2009.

[89] Lee Wang, Chuang Wang, Xing Xie, Josh Forman, Yansheng Lu, Wei-Ying Ma, and Ying Li. Detecting dominant locations from search queries. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 424–431. ACM, 2005.

[90] Riu Wang and Günter Neumann. Ontology-based query construction for geoclef. In *Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008*, volume 5706 of *Lecture Notes in Computer Science*, pages 880–884. Springer, 2009.

[91] Bo Yu and Guoray Cai. A query-aware document ranking method for geographic information retrieval. In *Proceedings of 4th InternationalWorkshop on Geographic Information Retrieval GIR 2007, ACM CIKM-2007*, pages 49–54, Lisbon, Portugal, 2007. ACM Press.

[92] Chengxiang Zhai. *Notes on the Lemur TFIDF Model*. School of Computer Science, Carnegie Mellon University, 2001. Technical report.

[93] Vivian Wei Zhang, Benjamin Rey, Eugene Stipp, and Rosie Jones. Geomodification in query rewriting. In *Proceedings of 3rd Workshop on Geographic Information Retrieval GIR'06*, Seattle, WA, USA, 2006. ACM Press. Extended abstract.