Expansión Automática de Peticiones en Búsqueda de Respuestas

por

María del Rosario Peralta Calvo
Ing., UTM

Tesis sometida como requisito parcial para obtener el grado de

Maestría en Ciencias en el área de Ciencias Computacionales

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica, INAOE
Febrero 2008
Tonantzintla, Puebla

Supervisada por:
Dr. Manuel Montes y Gómez
Investigador titular del INAOE

Dr. Luis Villaseñor Pineda
Investigador titular del INAOE

Resumen

Los sistemas de Búsqueda de Respuestas se enfocan en responder preguntas concretas formuladas por los usuarios. Para ello consideran tres procesos principales: el análisis de la pregunta, la recuperación de pasajes, y la extracción de la respuesta. Evidentemente todos estos procesos son necesarios, sin embargo la recuperación de pasajes mantiene un rol prioritario, pues sí esta etapa falla cualquier esfuerzo en la extracción de respuestas, por más sofisticado que sea, será infructuoso. En particular este trabajo de investigación se centra en la recuperación de pasajes. Básicamente propone un nuevo método para la expansión automática de preguntas. El método propuesto aplica técnicas de agrupamiento para separar los documentos relevantes y los no relevantes usados en la retroalimentación de relevancia automática, y considera una modificación a la técnica de Rocchio que permite emplear esquemas de recuperación boléanos. Los resultados experimentales muestran que aplicando el método de expansión de la consulta propuesto se obtiene un incremento aproximado de cuatro pasajes relevantes en promedio por pregunta, con respecto a los pasajes relevantes obtenidos previamente a la expansión.

Abstract

Question Answering systems allow answering natural language questions formulated by users. These systems mainly consider three processes: question analysis, passage retrieval and answer extraction. Evidently, all these processes are required, but the passage retrieval module has a key role. It is noticeable that if passage retrieval fails then any effort at subsequent modules will also fail independently of their suitability. In particular this research work focuses on the task of passage retrieval. It mainly proposes a new method for automatic question expansion. This method applies clustering techniques to separate the relevant and non relevant documents used for blind query expansion. It also considers an adaptation of the Rocchio approach in order to perform a Boolean retrieval. Experimental results show that the application of our method allows increasing the number of retrieved relevant passages per questions.

Agradecimientos

Mis más sinceros agradecimientos a mis asesores Dr. Manuel Montes y Gómez y Dr. Luis Villaseñor Pineda, quienes con su conocimiento, experiencia y amistad se logró concluir esta tesis.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo otorgado a través de la beca para estudios de maestría no. 201995.

Al Instituto Nacional de Astrofísica Óptica y Electrónica (INAOE) por las facilidades prestadas para la realización de mis estudios de maestría, en especial a la Coordinación de Ciencias Computacionales.

De forma muy especial a todos los integrantes del Laboratorio de Tecnologías de Lenguaje del INAOE, por compartir recursos y herramientas para el desarrollo de esta tesis.

A los Doctores Jesús Ariel Carrasco Ochoa y Aurelio López López por su confianza y apoyo al decidir darme la oportunidad de estudiar la maestría.

A mis sinodales: Dra. Angélica Muñoz Meléndez, Dr. Eduardo Francisco Morales Manzanares y Dr. Aurelio López López, por su tiempo y aportaciones a la tesis.

A mis hermanos: Blas $^{+}$, Antonio, César, Abdón y Fidel, por su motivación y apoyo en todo momento.

A mis sobrinos: Antonio, Fernanda, Hilda, Arturo, Fernando, Merari, Miguel, Mariana, Ramón, Wendy y al bebe que aún no se que será pero que ya cuenta, por ser la alegría de mi vida, que sigan adelante y logren al máximo sus metas.

A Israel, una vez más, por su cariño, apoyo y motivación, gracias de todo corazón.

Al M.C. Mario Alberto Moreno Rocha por impulsarme en todo momento para continuar con mis estudios de maestría.

A mis primas: Gaby y Adriana por sus porras y oraciones en todo momento difícil de la maestría.

A mis primos: Charbel, Antonio y Judith para que les sirva de motivación en sus estudios.

A mi tío Teofilo y su familia, por su hospitalidad y cariño que fue fundamental desde el inicio hasta el final de la tesis.

A mi familia en general por sus palabras de aliento y oraciones que contribuyeron enormemente para que lograra mi maestría, muy en especial a mis tíos, tías, demás primos, primas, cuñadas y cuñado.

A Iris, Adrián, Argelia que a pesar de la distancia no dejaron de motivarme como verdaderos amigos.

A Nadia, Rosy, Sayde, Coral, Erika, Esteban, Gershom por su amistad y compañerismo en estos dos años de la maestría.

A mis compañeros de la maestría porque a pesar de las circunstancias logramos la meta.

Dedicatoria

A **Dios**,
por darme la salud y la vida para cumplir este reto,
por cuidar a la gente que amo y
con la que ahora puedo compartir esta alegría.

A mis padres: **Yolanda y Fernando**,
por su amor, apoyo incondicional, confianza y comprensión.

A mi hermana **Yolanda**,
por su incansable apoyo cuando más lo necesite.

Contenido

Lista de Figuras

Lista de Tablas

Capítulo 1

Introducción

Con el desarrollo de diferentes tecnologías como la Internet, el acceso a grandes colecciones de documentos es cada vez más sencillo. Aún los libros, las revistas y otros medios de consulta se encuentran ahora convertidos en recursos electrónicos que pueden accederse a través de la Internet. El tamaño de las colecciones que ahora podemos almacenar es enorme dificultando encontrar un determinado fragmento de información requerido por un usuario. De ahí que diferentes líneas de investigación dentro del Tratamiento Automático de Textos proponen soluciones diversas al problema de búsqueda de información. Dos áreas son de particular interés en el contexto de esta tesis: la Recuperación de Información (RI) y la Búsqueda de Respuestas (BR).

La RI presenta al usuario una lista de documentos ordenados por su relevancia respecto a cierta petición. Para ello, se han planteado diferentes técnicas para identificar aquellos documentos en que aparecen los términos de la petición, para lo cual es importante calcular la relevancia de los términos en el proceso de discriminación (Llopis et al., 2002). Una vez calculada la lista de documentos relevantes, el usuario determina si la información contenida en alguno de estos responde a sus necesidades. En el área de BR se intenta proporcionar al usuario una respuesta concisa a una

pregunta formulada por el usuario en lenguaje natural, como un ejemplo de preguntas tenemos aquellas que se refieren al nombre de un personaje en algún acontecimiento importante, por lugares o fechas los cuales están asociados a eventos, entre otros.

1.1 Descripción del problema

La problemática que aborda el presente trabajo consiste en la recuperación de información dentro de la tarea de la búsqueda de respuestas (BR). Tradicionalmente, un sistema de BR está formado por: un módulo de análisis de la pregunta, un modulo de recuperación de pasajes, y un módulo de extracción de la respuesta. El primer módulo determina el tipo de respuesta que se espera, por ejemplo, si es una persona, lugar, fecha o cantidad. El segundo recupera aquellos fragmentos de texto, o *pasajes*, que son relevantes a la pregunta y donde se presume se encontrará la respuesta. Por último, el tercer módulo trata de extraer la respuesta exacta a partir de los pasajes recuperados (Sanchis et al., 2007).

Para tener éxito en la extracción de la respuesta correcta es muy importante trabajar con fragmentos de textos adecuados, aquellos que incluyen la posible respuesta. De ahí la gran importancia del módulo de recuperación de pasajes. La Recuperación de Pasajes (RP) puede verse como un filtro que reduce la colección original de documentos a un conjunto de pasajes de los cuales la respuesta será extraída. El objetivo es eliminar tantos fragmentos irrelevantes como sea posible y concentrar aquellos que contienen la respuesta. Así, el desempeño de los sistemas de RP determina un límite para los sistemas de BR (Sanchis et al., 2007).

Existen diversos trabajos que buscan mejorar el desempeño de los sistemas de recuperación de pasajes. Un enfoque en particular es de interés para nuestro trabajo, al considerar el sistema de RP como una caja negra y trabajar sobre los pasajes que se recuperan, así como la petición que recibe

el sistema de RP de manera totalmente independiente. Lo anterior permite analizar la relevancia de dichos pasajes así como la pertinencia de la petición sin considerar los detalles de la técnica empleada dentro del sistema de RP.

La intención final es manipular la petición original para mejorar el desempeño en la recuperación de información. Diferentes métodos se han propuesto en RI, uno de estos es conocido como expansión de la petición. Ésta consiste en agregar términos a la petición original para eliminar posibles ambigüedades y mejorar el desempeño en la recuperación. A continuación se presenta la solución que se exploró en esta tesis para determinar con qué términos expandir la petición.

1.2 Solución propuesta

La *expansión de la petición* es un método para mejorar la recuperación de información, que consiste básicamente en reformular la petición original, para construir una nueva petición y generar nuevos resultados. En RI una de las técnicas que se ha aplicado para la expansión de la petición es la *retroalimentación de relevancia,* cuya idea consiste en mejorar la petición con información previamente recuperada, ya sea de forma manual o automática. La primera forma consiste en involucrar al usuario, al permitirle elegir la información (términos o documentos) para hacer la expansión. La segunda forma también conocida como *expansión ciega* consiste en determinar automáticamente los términos con los cuales se expande la petición. En la *expansión ciega* se identifican aquellos términos –supuestamente asociados a los términos de la petición- a través de los documentos recuperados con la petición original. También, en la *expansión ciega* se parte de que el sistema de recuperación es capaz de entregar los documentos por orden de relevancia, y que aquellos con mayor relevancia son pertinentes a la petición, para que una vez seleccionados los documentos con mayor relevancia se

calcule la ocurrencia de los términos. Aquellos términos más frecuentes –diferentes a los de la petición inicial– son considerados para la expansión.

Este tipo de técnicas han generado resultados favorables en RI. Lo anterior no puede trasladarse fácilmente a los sistemas de BR, debido a que en estos sistemas se parte de una pregunta formulada en lenguaje natural, por lo que el proceso de la expansión de la petición alteraría de forma considerable la estructura de la pregunta. Por otro lado, en RI el cálculo de ocurrencias de términos se realiza a nivel de documentos y en BR se realiza a nivel de pasajes, lo que complica la identificación de los términos a utilizar en la expansión.

El presente trabajo aplica la expansión ciega de peticiones en el contexto de búsqueda de respuestas, con la intención de medir el impacto de este proceso en la recuperación de pasajes. El trabajo propone dos esquemas para seleccionar el conjunto de pasajes relevantes, a partir de los cuales se identifica los términos para reformular la petición.

1.3 Objetivos

Esta tesis tiene los siguientes objetivos:

Objetivo general: Proponer y aplicar un método de expansión automática de la petición para la recuperación de pasajes en un sistema de búsqueda de respuestas.

Objetivos particulares:

- Adaptar y evaluar las técnicas tradicionales de retroalimentación de relevancia en la expansión automática de peticiones, para la recuperación de pasajes en un sistema de búsqueda de respuestas.
- Proponer y aplicar un método basado en agrupamiento para la expansión de peticiones orientado a búsqueda de respuestas.

1.4 Organización de la tesis

La tesis se organiza de la siguiente manera: en el capítulo 2 se da a conocer el estado del arte de esta tesis. El capítulo 3 describe el método propuesto explicando sus puntos importantes. El capítulo 4 detalla los experimentos y resultados para la evaluación del método. Por último, en el capítulo 5 se exponen las conclusiones de la investigación desarrollada y el trabajo futuro de dicha investigación.

Capítulo 2

Estado del Arte

En el presente capítulo se da a conocer el estado del arte para ubicar al lector en el área donde se enfoca esta tesis.

2.1 Búsqueda de Respuestas

Como ya se ha mencionado, en las últimas décadas, el aumento de la información que circula a través de la Internet ha dificultado a los usuarios el acceso a los documentos. Los sistemas de Recuperación de Información (RI), como los motores de búsqueda son los que comúnmente en la Internet se encargan de que los usuarios realicen sus peticiones y permitan el acceso a la información que deseen. Estos sistemas de RI seleccionan y recuperan un conjunto de documentos a partir de las necesidades de información de los usuarios. También se ha mencionado, la exigencia de los usuarios para obtener información concreta a sus peticiones, por lo que surgen dentro del área de Tratamiento Automático de Textos, los sistemas de Búsqueda de Respuestas (BR) cuyo objetivo es proporcionar la respuesta correcta a la pregunta formulada por el usuario en lenguaje natural (Ferrández, 2004).

Para lograr dicho objetivo, la arquitectura general de un sistema de BR consiste de tres módulos principales: el procesamiento de la pregunta, la

recuperación de pasajes y la extracción de la respuesta. La figura 2.1 muestra de forma gráfica dicha arquitectura (Harabagiu y Moldovan, 2003).

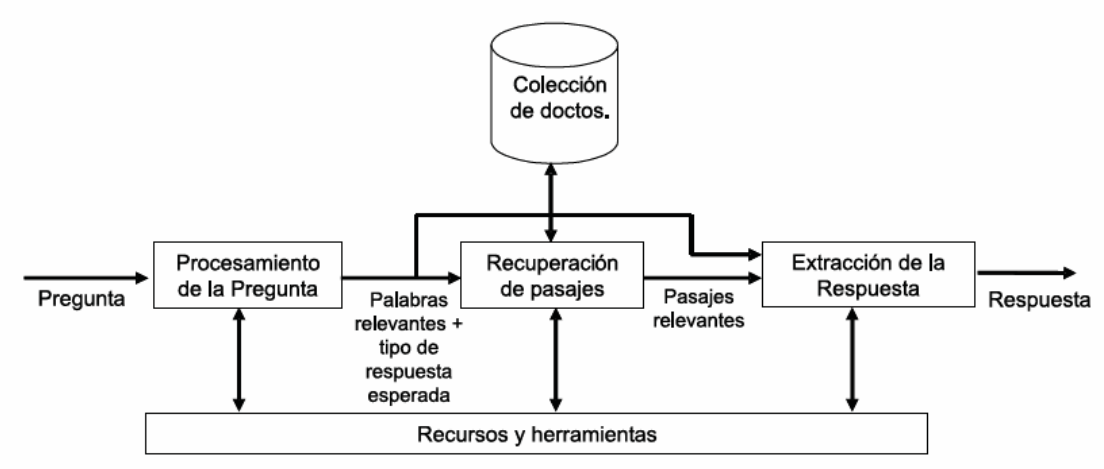

Figura 2.1 Arquitectura de un sistema de BR

La función del módulo del **procesamiento de la pregunta** consiste en realizar un análisis exhaustivo de la pregunta, generando información que permitirá localizar la respuesta correcta. Este módulo trata de identificar el tipo semántico de la respuesta esperada (un lugar, una persona, una expresión de tiempo o cantidad, etc.) o aquellos términos de la pregunta que puedan ayudar a localizar los documentos que pueden contener la información buscada.

El módulo de **recuperación de pasajes** tiene como función proporcionar un conjunto de pasajes que presumiblemente pueden contener la respuesta a la pregunta. Este módulo parte previamente de una recuperación de documentos que reduce una colección de millones de documentos a una selección de algunos cientos de documentos, sobre los que se realizará una búsqueda más minuciosa entre su contenido. Estos cientos de documentos junto con lo que resulte del análisis de la pregunta (convertida en una petición) son la entrada para la recuperación de pasajes.

Debido a que los documentos pueden ser de un tamaño considerable, se restringe el espacio sobre el cual se realizará la búsqueda, pasando de documentos a párrafos o fragmentos contiguos de texto que pueden contener la información requerida por el usuario.

Finalmente, el módulo de **extracción de la respuesta** tiene como función tomar la información que haya resultado de la recuperación de pasajes para extraer la respuesta.

Evidentemente todos los componentes son necesarios y relevantes para lograr contestar la pregunta. Sin embargo, tal como lo señalan algunas investigaciones (Tellex et al., 2003; Sanchis et al., 2007), la recuperación de pasajes es de vital importancia puesto que de los pasajes que se recuperen se extraerá la respuesta a la pregunta.

En esta tesis nos enfocaremos en la recuperación de pasajes, por ello en la siguiente sección se tratará este módulo.

2.2 Recuperación de pasajes en BR

La Recuperación de Pasajes es la fase previa a la extracción de la respuesta y tiene como objetivo proporcionar un conjunto de pasajes que presumiblemente pueden contener la respuesta a la pregunta. Estos pasajes recuperados son extraídos de una colección de documentos y contienen información relevante a la pregunta.

En esta recuperación de pasajes se aplican técnicas como en RI para recuperar documentos, sólo que este módulo se enfrenta a dos problemas:

a) Cómo dividir un documento en pasajes.

b) Cómo evaluar la relevancia de cada pasaje, considerando que son textos cortos.

Existen tres formas de dividir un documento en pasajes: modelos de discurso, modelos semánticos y modelos de ventana. El basado en el discurso depende de un discurso textual que usa la propiedad estructural de los documentos, tales como oraciones, párrafos o secciones. El basado en la semántica divide cada documento en piezas semánticas según los diferentes temas en el documento. Por último, el basado en ventanas simplemente considera un número de palabras para determinar los límites de los pasajes (Callan, 1994; Llopis et al., 2002; Sanchis et al., 2007).

Existen dos métodos para evaluar la relevancia de cada pasaje con respecto a la pregunta que ha realizado el usuario (la pregunta en este proceso de RP se toma como una petición) (Tellex et al., 2003; Sanchis et al., 2007):

a) El método basado en traslape que considera la coincidencia de los términos de la petición y los términos de los pasajes.
b) El método basado en densidad que considera la distribución de los términos de la petición en los términos de los pasajes.

Tellex et al. (2006) comparan en su investigación distintos algoritmos de RP, de los cuales concluyen que aquellos basados en densidad son los mejores, ya que no sólo consideran que los términos de la petición coincidan en el pasaje, sino que asigna una puntuación al pasaje cuando verifica que tan cerca están –unos de otros– los términos claves de la petición en el pasaje.

Un sistema de recuperación de pasajes debe ser sometido a evaluaciones, Roberts y Gaizauskas (2004) proponen dos métricas para ser usadas bajo el contexto de BR, que son la cobertura y la redundancia y las definen como sigue: Sea Q el conjunto de las preguntas, P la colección de pasajes, $A_{P,q}$ el subconjunto de P el cual contiene respuestas correctas para $q \in Q$, y $R^{S}_{P,\,q,\,n}$ representa los primeros n pasajes en P recuperados por un sistema de recuperación S dada una pregunta q.

La **cobertura** de un sistema de recuperación *S* para un conjunto de preguntas *Q* y una colección de pasajes *P* a un rango *n* queda denotada por la expresión (2.1.1):

$$cobertura^{S}(Q,P,n) = \frac{\left|\left\{q \in Q \middle| R^{S}_{P,q,n} \cap A_{P,q} \neq \phi\right\}\right|}{\left|Q\right|} \qquad (2.1.1)$$

Esta cobertura da la proporción del conjunto de preguntas para la cual al menos una respuesta correcta puede ser encontrada en los primeros *n* pasajes recuperados para cada pregunta.

La **redundancia** en la respuesta de un sistema de recuperación *S* para un conjunto de preguntas *Q* y colección de pasajes *P* a un rango *n* queda denotada por la expresión (2.1.2):

$$redundancia^{S}(Q,P,n) = \frac{\sum_{\forall q \in Q}\left|R^{S}_{P,q,n} \cap A_{P,q}\right|}{\left|\left\{q \in Q \middle| R^{S}_{P,q,n} \cap A_{P,q} \neq \phi\right\}\right|} \qquad (2.1.2)$$

Esta redundancia da el número promedio por pregunta de pasajes dentro de los primeros *n* pasajes recuperados que contengan la respuesta correcta.

Las evaluaciones en los sistemas de RP permiten conocer el desempeño de estos en el proceso de recuperación de información relevante. Estos sistemas de recuperación de información sobre pasajes están en constante evolución, ya que de la información que se genere depende la extracción de la respuesta.

En la siguiente sección se detalla la expansión de peticiones que ha sido usada para mejorar la recuperación de información.

2.3 Expansión de peticiones

Al realizar una búsqueda a través de un sistema de RI, el usuario expresa su petición haciendo uso de una serie de palabras. Estos sistemas buscan establecer una relación entre la petición y los documentos de la colección, recuperando sólo aquellos que coincidan con la petición. Un problema muy común en RI es que existen muchos documentos con los cuales pueden darse coincidencias con uno o más términos de la petición, y no tener alguna relación con el tema, porque tal vez sean palabras similares o ambiguas, y generan documentos no relevantes al usuario.

Distintas investigaciones se han enfocado a mejorar la efectividad de dichos sistemas de RI, un método que ha sido investigado ampliamente para mejorar el desempeño de los sistemas de RI, se conoce como expansión de la petición (Billerbeck y Zobel, 2004; Hsu et al., 2006; Cui et al., 2002).

El proceso de expansión de la petición consiste en reformular la petición original para mejorar el desempeño de la recuperación de información. En esta tesis, la reformulación consiste en agregar y eliminar términos para modificar la petición original.

Klink (2001) indica que el principal reto al que se enfrenta la expansión de peticiones es la pregunta: ¿Cuáles términos deberían ser incluidos en la formulación de la petición?

Para responder a dicha pregunta, la expansión de peticiones se ha dado bajo dos enfoques.

1. El primero consiste en usar términos relacionados a la petición según algún recurso lingüístico de referencia (ontología o tesauros).
2. El segundo consiste en usar los términos que más ocurren en los documentos superiores del ordenamiento después de procesar la petición original (retroalimentación de relevancia).

2.3.1 Recursos lingüísticos

El primer enfoque de la expansión de la petición consiste en reformular la petición con información que resulte de la relación taxonómica (sinónimos o hipónimos) entre los términos de la petición y los términos de los documentos de la colección, recurriendo a algún recurso lingüístico como WordNet o tesauros.

Mandala et al. (1999) exponen que la expansión de la petición basada en un tesauro especializado tiene éxito si sólo se realizan peticiones que tengan relación con dicho recurso, por lo que la expansión se limita al uso de los términos que en el recurso lingüístico se encuentren y que se relacionen con los términos de la petición original.

El trabajo de Gelbukh (2000) consiste en expandir la petición haciendo uso de información que se extrae de un tesauro. Es decir, la petición realizada por el usuario se sustituye automáticamente por una expresión lógica que involucre los términos extraídos del tesauro. En dicho trabajo, la forma de relacionar los términos de la petición original con los que se encuentren en el tesauro es usando su morfología, sinónimos, hipónimos y hiperónimos.

En el trabajo de Navigli y Velardi (2003) se realiza la expansión de peticiones haciendo uso de la desambiguación del sentido de las palabras (WSD sus siglas en inglés). La nueva petición se forma con palabras que se extraigan de la red semántica y que se relacionen con los términos de la petición original. Dicha red se crea considerando la relación taxonómica de sinónimos e hiperónimos de las palabras. WordNet es el recurso lingüístico de donde se extrae el sentido de las palabras y las relaciones taxonómicas.

2.3.2 Retroalimentación de relevancia

Para el segundo enfoque de la expansión de la petición se tiene la técnica de retroalimentación de relevancia que por muchos años se ha sugerido como la solución para modificar la petición (Harman, 1992). Es un mecanismo para refinar el proceso de búsqueda usando conocimiento obtenido en la búsqueda inicial para una búsqueda final (Billerbeck y Zobel, 2004). La retroalimentación de relevancia usualmente se lleva acabo de la siguiente manera (Noguera y Llopis, 2007):

- Se ejecuta una búsqueda usando la petición original
- Se seleccionan *t* términos de un subconjunto de los documentos recuperados
- Se agregan *t* términos a la petición original para formular la nueva petición.
- Se ejecuta la nueva petición generando una nueva lista de documentos recuperados.

La retroalimentación de relevancia tiene que resolver de qué forma hacer la selección de los documentos de los cuales se estarían tomando los términos para la expansión de la petición. Para esto surgen dos maneras de realizar la retroalimentación de relevancia:

1. Manual, cuando el usuario influye en la retroalimentación al indicar qué documentos son relevantes;
2. Automática (o ciega), cuando se consideran como relevantes los *n* primeros documentos recuperados. Por otro lado, es posible considerar los restantes documentos (más allá del umbral *n*) como no relevantes o, en ocasiones, sólo un subconjunto de ellos. (Grossman y

Frieder, 2004). En este caso, se confía plenamente en el ordenamiento resultante del proceso de recuperación de documentos.

Chang y Hsu (1998) aplican una retroalimentación manual para la expansión de la petición. Con base en el agrupamiento de documentos, el usuario indica (basándose en conceptos) qué tópicos del agrupamiento se consideran para la retroalimentación y expansión de la petición, con términos extraídos de la selección hecha por el usuario.

Xu y Croft (2000) explican la técnica de análisis de contexto local (ACL). Esta técnica utiliza *conceptos* para la expansión. Xu y Croft definen *concepto* como un sustantivo. El análisis de contexto local asigna un valor entre 0 y 1 a los conceptos, ordenándolos de mayor a menor, estableciendo así un rango. Dicho valor se asigna según el número de co-ocurrencias que tengan los conceptos con respecto a los términos de la petición, en los primeros documentos ordenados. El rango ya establecido entre conceptos, permite que para la expansión se consideren sólo aquellos conceptos que tengan un valor alto en el rango.

Existen distintas técnicas dentro de la retroalimentación de relevancia que se aplican para el proceso de expansión de la petición, una de las más representativas es la técnica de Rocchio que en la siguiente sección se detalla.

2.3.3 Técnica de Rocchio

La técnica de Rocchio consiste en medir la similitud entre el vector de la petición Q_0 y el vector del documento D_i extraído de una colección de documentos D. Para esto se asume que el usuario ha realizado una petición de donde obtiene un conjunto de documentos. Se consideran dos grupos de documentos, los relevantes y los no relevantes. Entonces se tiene que el conjunto R contiene n vectores de documentos relevantes y el conjunto S

contiene *m* vectores de documentos no relevantes (Harman, 1992). La técnica de Rocchio construye la nueva petición Q_1 a partir de la anterior Q_0 usando la expresión (2.1.3) (Grossman y Frieder, 2004):

$$Q_1 = Q_0 + \beta \sum_{k=1}^{n} \frac{R_k}{n} - \gamma \sum_{k=1}^{m} \frac{S_k}{m} \qquad (2.1.3)$$

Donde:

Q_1 es el nuevo vector petición,

Q_0 es el vector de la petición inicial,

R_k es el vector del *k* documento relevante,

S_k es el vector del *k* documento no relevante,

n es el número de documentos relevantes,

m es el número de documentos no relevantes,

β y γ son los parámetros que controlan la contribución relativa de documentos relevantes y no relevantes (coeficientes de Rocchio).

La expresión (2.1.3) indica que los vectores de documentos que resulten relevantes son sumados a la petición inicial, y los vectores de los no relevantes se substraen. Se asegura que la nueva información no anula completamente a la petición original. Todos los vectores modificados son promediados con respecto al número de documentos relevantes y no relevantes. Harman (1992) menciona que $\beta = 0.75$ y $\gamma = 0.25$, son los valores con los cuales la expresión (2.1.3) ha tenido mejores resultados. La técnica de Rocchio pretende hallar más documentos relevantes que tengan entre sus términos aquellos que coincidan con los de la petición original. A partir de la técnica anterior se derivan otras dos:

- **Ide regular** (expresión 2.1.4) que consiste en hacer β = γ = 1, y no forman vectores promedio de los documentos relevantes y no relevantes para formar la nueva petición.
- **Ide dec-hi** (expresión 2.1.5) que calcula los términos a agregar (o eliminar) dando mayor importancia al contexto local de cada documento.

En seguida se muestran las expresiones para estas técnicas (Harman, 1992).

$$Q_1 = Q_0 + \sum_{k=1}^{n} R_k - \sum_{k=1}^{m} S_k \qquad (2.1.4)$$

$$Q_1 = Q_0 + \sum_{k=1}^{n} (R_k - S_k) \qquad (2.1.5)$$

Con respecto al pesado de los términos, estas técnicas re-pesan los términos de la petición sumando los pesos de ocurrencia actual de estos términos de la petición en los documentos relevantes, y restando los pesos de estos términos que ocurren en los documentos no relevantes.

Las peticiones son expandidas agregando los términos no incluidos en la petición original que están en los documentos relevantes y que no están en los no relevantes. También se expanden usando pesos positivos y negativos, si los términos provienen de documentos relevantes o no relevantes, respectivamente.

2.4 Expansión de peticiones en BR

La expansión de peticiones ha comenzado a incursionar en el área de BR. Usualmente forman las nuevas peticiones basándose en los siguientes enfoques (Greenwood, 2004):

1) Con palabras o conceptos relacionados a la pregunta (sinónimos o variantes morfológicas).
2) Con términos parecidos que co-ocurren con instancias del tipo de respuesta esperada.

Para el primer enfoque, Bilotti et al. (2004) construyen la expansión de petición agregando los términos de la petición con algunas de sus variantes morfológicas para recuperar información. Por ejemplo, dada la pregunta *"What lays blue eggs?"*, la petición expandida se expresa: *blue* $\wedge$ *(eggs* $\vee$ *egg)* $\wedge$ *(lays* $\vee$ *laying* $\vee$ *lay* $\vee$ *laid)*.

También en el primer enfoque, Greenwood (2004) explora un método para formular la expansión de preguntas, donde éstas se refieran a información relacionada a una localidad o lugar. Las preguntas son expandidas usando relaciones correspondientes a una localización extraída de WordNet. Por ejemplo, si en la pregunta se desea saber acerca de un país, el método extrae de WordNet la nacionalidad o el idioma del país y lo ubica en la expansión. Es decir, para la pregunta: "*What is the capital of Syria?*", la expansión sería: *capit*[1] *(syria syrian)*.

Negri (2004) combina parte de los dos enfoques, presenta un método para expandir la petición en el contexto de BR, basado en retroalimentación de relevancia para ayudar a la desambiguación del sentido de las palabras en la petición. Dicho método de expansión consiste en utilizar las palabras claves de la pregunta para recuperar los primeros *n* documentos, y sobre

[1] Término que utiliza Greenwood (2004) para expresar la relación "capital".

estos elegir los más frecuentes sentidos de las palabras (extraídos de WordNet) de los términos de la pregunta. Estos sentidos de las palabras son los que se consideran para la expansión en lugar de optar por las palabras más relevantes que aparecen en estos *n* documentos recuperados. En consecuencia, la petición expandida se realiza agregando los términos semánticamente relacionados a dichos sentidos de las palabras.

Aplicando el segundo enfoque, Monz (2003) investigó el efecto de expandir la petición para cierto tipo de preguntas: aquellas que respondan a medidas como altura, longitud, edad, entre otras. Esta selección se debe a que las respuestas para estas preguntas requieren contener una unidad de medida, lo cual las limita. La expansión no sólo se realizaba agregando términos, sino que se agrupaban términos como una forma de establecer términos alternativos que se derivaban de uno solo. Por ejemplo, de la pregunta *"How high is Mount Kinabalu?"*, se define su tipo, *Question type: number-height*, y se realiza la expansión con esta petición, *Query: mount kinabalu alt (meter,inch,foot,centimeter).*

La mayoría de estas investigaciones han hecho uso de recursos externos para relacionar los términos, pero no se descarta considerar las co-ocurrencias de dichos términos para construir la petición. Esto ha generado resultados favorables que se pueden considerar para mejorar la recuperación de información en el área de BR.

Cabe mencionar que para ampliarse a otros idiomas, estos métodos de expansión requieren no depender de recursos los cuales sólo consideran un idioma, tal es el caso de WordNet, que está formado sólo para el idioma Inglés. Una alternativa para la independencia del idioma puede ser el uso de las co-ocurrencias de términos en la información recuperada, para extraer los términos que serán usados en la expansión de la petición. El uso de este segundo enfoque de la expansión, indica que se está aplicando la retroalimentación de relevancia que utiliza como conocimiento relevante la información que se genera al ejecutar la petición original.

En particular para el idioma Español no ha sido explorada la retroalimentación de relevancia en la expansión de peticiones, por lo que éste es el objetivo de este trabajo de tesis en el área de BR.

Capítulo 3

Método de Expansión

La expansión de peticiones en RI ha ido evolucionando a través de dos enfoques: el uso de recursos externos y el uso de la técnica de retroalimentación de relevancia. La retroalimentación de relevancia muestra ciertas ventajas con respecto al uso de recursos externos como son: la independencia del idioma y del dominio de la información.

El proceso de expansión de la petición no ha sido ajeno para el área de BR. Este proceso permite mejorar la recuperación de información para que sea cada vez más información relevante la que se logre recuperar y aporte beneficios a la recuperación de pasajes en los sistemas de BR.

Debido a que la recuperación de pasajes es un proceso interno en la arquitectura general de un sistema de BR, la retroalimentación de relevancia en su forma automática se ajusta para aplicarse dentro de estos sistemas. Es decir, se considera sólo la información de los pasajes recuperados para la expansión de la petición.

La forma tradicional en como se ha aplicado la retroalimentación automática consiste en seleccionar los primeros *n* pasajes relevantes y considerar el resto como no relevantes, es decir, separamos los pasajes recuperados en dos grupos. Por lo tanto, esta tesis se ha orientado en hacer uso de la retroalimentación de relevancia automática tanto en su forma

tradicional (definiendo el umbral *n* arbitrariamente); y usando agrupamiento de pasajes. Dicho agrupamiento permite reunir aquellos pasajes similares y establecer un juicio más puntual de su relevancia, para la expansión automática de peticiones en BR. Así pues, en la siguiente sección se presenta el método propuesto y se describe cada uno de los módulos que lo conforman.

3.1 Método propuesto

El método propuesto de expansión automática de la petición tiene como objetivo apoyar al módulo de recuperación de pasajes, para contar con pasajes relevantes de donde se pueda extraer la respuesta. Esta recuperación de pasajes consiste en tomar como petición al conjunto de términos que se deriva de la pregunta, para procesar dicha petición y recuperar un cierto número de pasajes. A estos pasajes recuperados se les aplican el método propuesto, para lograr la expansión de la petición. La figura 3.1 indica los componentes de dicho método.

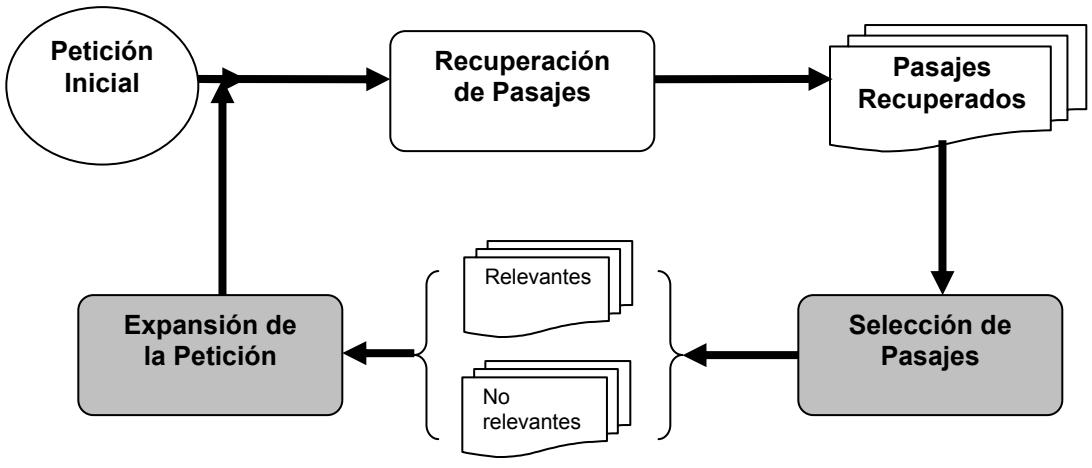

Figura 3.1 Método propuesto

Selección de pasajes es un proceso cuya idea es establecer dos grupos de pasajes: los relevantes y los no relevantes. Se aplican dos criterios para realizar este proceso: uno consiste en establecer un umbral y el otro consiste en aplicar una técnica de agrupamiento.

Expansión de la petición es un proceso que consiste en aplicar la retroalimentación de relevancia a través de la técnica de Rocchio para construir la nueva petición, tomando en cuenta la información de pasajes relevantes y no relevantes.

3.2 Seleccionar pasajes

El objetivo de este módulo, como ya se había mencionado, consiste en dividir los pasajes recuperados en relevantes y no relevantes. Los dos criterios con los que se trabajó en esta tesis se detallan enseguida.

El primer criterio sigue el enfoque tradicional. En el cual se acepta como correcto el ordenamiento de los pasajes recuperados; y así se asumen como relevantes los primeros *n* pasajes recuperados. Cabe mencionar que se corre un riesgo con este método, ya que la información relevante puede quedar fuera del rango del umbral (*n*) y considerarse como información no relevante o viceversa.

El segundo criterio consiste en determinar el conjunto de pasajes relevantes a través de un proceso de agrupamiento automático. En la siguiente sección se detalla este proceso.

3.2.1 Agrupamientos de pasajes

El agrupamiento o *clustering* es un método del aprendizaje no supervisado, donde se cuenta con datos no etiquetados a los cuales habría que

agruparlos, de tal forma que los objetos de un grupo tengan una similitud alta entre ellos y baja con los objetos de otros grupos.

Con el agrupamiento se pretende tener grupos asociados, es decir, que los pasajes estén estrechamente relacionados con base en la similitud entre los términos que los conforman.

De forma general, la tarea de agrupamiento de pasajes involucra los siguientes pasos:

1. Transformar los pasajes en una representación apropiada para su comparación.
2. Comparar los pasajes y construir una matriz considerando la ocurrencia de los términos contenidos en los pasajes.
3. Aplicar un algoritmo de agrupamiento y obtener los grupos de pasajes.

La figura 3.2 indica, de manera gráfica la forma en cómo se hizo el proceso de agrupamiento de pasajes.

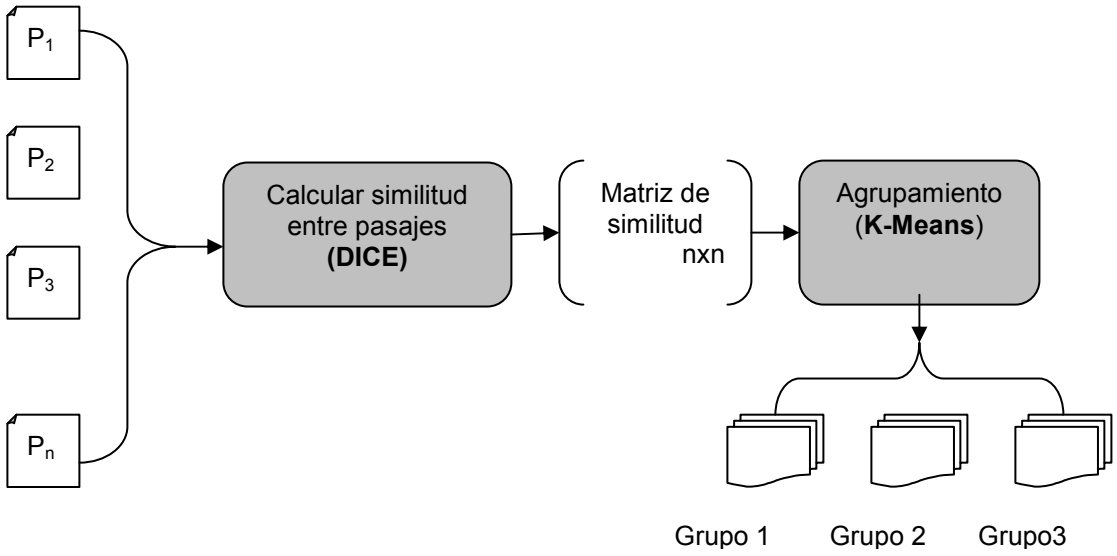

Figura 3.2 Agrupamiento de pasajes

A continuación se detallan cada uno de los procesos que se realizan en el agrupamiento de pasajes dentro del método propuesto.

Para **transformar los pasajes** en una representación apropiada se considera la pregunta formulada en lenguaje natural, la petición resultante del análisis de la pregunta, y uno de los pasajes recuperados para dicha petición. La tabla 3.1 muestra un ejemplo de este tipo.

Tabla 3.1 Ejemplo de una pregunta, petición y pasaje recuperado

Pregunta	**Petición**	**Pasaje**
¿Qué volcán entró en erupción en junio de 1991?	volcan entro erupcion junio 1991	2485660 0.5507619 EFE19950127-15641 ECUADOR-GALAPAGOS ENTRA EN ERUPCION VOLCAN EN UNA DE LAS ISLAS DEL ARCHIPIELAGO Quito , 27 ene (EFE).- El volcán Fernandina , de la isla del mismo nombre del archipiélago ecuatoriano de las Galápagos , entró en actividad el miércoles pasado en un sector no habitado , informaron hoy , viernes , fuentes gubernamentales .

El pasaje presenta un formato específico que consiste en un identificador de pasaje, un peso de relevancia, el identificador del encabezado del documento de donde se extrajo, y el pasaje extraído.

La transformación del pasaje consiste en realizar primero una limpieza de los pasajes, esto es eliminar información no relevante, como son los identificadores del pasaje, el peso de relevancia, el identificador del encabezado del documento y las palabras vacías. Como segundo paso se convierten a un mismo formato las palabras que conforman el pasaje, es decir, debido a que puede haber mayúsculas y minúsculas, se elige el formato de minúsculas, y se eliminan los términos repetidos, para que posteriormente esto permita formar el vocabulario. La tabla 3.2 muestra el ejemplo de un pasaje original y el resultado del pasaje al ser transformado.

Tabla 3.2 Ejemplo de un pasaje trasformado

Pasaje original	**Pasaje transformado**
2485660 0.5507619 EFE19950127-15641 ECUADOR-GALAPAGOS ENTRA EN ERUPCION VOLCAN EN UNA DE LAS ISLAS DEL ARCHIPIELAGO Quito , 27 ene (EFE).- El volcán Fernandina , de la isla del mismo nombre del archipiélago ecuatoriano de las Galápagos , entró en actividad el miércoles pasado en un sector no habitado , informaron hoy , viernes , fuentes gubernamentales .	ecuador galapagos entra erupcion volcan islas archipielago quito 27 enero fernandina isla ecuatoriano entro actividad miercoles pasado sector habitado informaron viernes fuentes gubernamentales

Para **comparar los pasajes** se calcula la similitud con el coeficiente de Dice. Éste nos proporciona un valor numérico, que nos indicará qué tan similar es un pasaje a otro pasaje. El coeficiente de Dice es denotado por la expresión (3.1.1) (Grossman y Frieder, 2004).

$$C\_Dice(P_i, P_j) = \frac{2\sum_{k=1}^{t} w_{ik} w_{jk}}{\sum_{k=1}^{t} (w_{jk})^2 + \sum_{k=1}^{t} (w_{ik})^2} \qquad (3.1.1)$$

Donde:

k va de 1 al número total de términos del vocabulario *t*,

w_{ik} es igual a 1 si el término *k* ocurre en el pasaje P_i y 0 en caso contrario,

w_{jk} es igual a 1 si el término *k* ocurre en el pasaje P_j y 0 en caso contrario.

Aplicando esta medida de similitud se va construyendo una matriz de *nxn*, donde cada valor en la celda de la matriz representa el valor que arroje el coeficiente de Dice al compararse el pasaje P_i con el P_j.

Esta matriz de similitud en consecuencia es el dato de entrada para aplicar al algoritmo de agrupamiento. Para esta tesis se eligió el algoritmo **K-Means**, que se muestra en seguida:

- *Entradas*: conjunto de documentos, *k*.
    1. Seleccionar *k* documentos como centros de grupos.
    2. Asignar cada documento al grupo más cercano.
    3. Recalcular los centros de los grupos (punto medio).
    4. Si el criterio de convergencia no se alcanza, entonces regresar al paso 2
- *Salidas*: *k* grupos de documentos.

Se optó por usar **K-Means** por las ventajas que proporciona al trabajar con datos numéricos representados a través de la matriz de similitud.

El agrupamiento de pasajes generará grupos de pasajes sobre los cuales habrá que realizar la selección del grupo relevante, siendo éste el objetivo principal del proceso de seleccionar pasajes. Dicha selección se explica en la siguiente sección.

3.2.2 Selección del grupo relevante

La selección del grupo relevante tiene como objetivo identificar de entre los grupos generados en el paso anterior el grupo relevante.

La selección se basa en la idea de que el grupo con mayor presencia de los términos de la petición será el más relevante. Para ello se calculan dos valores: la frecuencia relativa simple y la frecuencia relativa acumulada. Para calcular la **frecuencia relativa simple** se mide la frecuencia de cada término de la petición (sin palabras vacías) en los pasajes contenidos en cada grupo, estas frecuencias se normalizan con respecto al número de pasajes de cada grupo. Para calcular la **frecuencia relativa acumulada** se suman todas las frecuencias relativas simples calculadas en cada grupo.

Con esta información es posible determinar el o los grupos relevantes. Para esto se proponen dos criterios:

- El primer criterio consiste en elegir un sólo grupo, aquel con la mayor frecuencia relativa acumulada. A este grupo relevante formado por un sólo grupo lo denominaremos **grupo relevante individual**.
- El segundo criterio consiste en determinar como relevantes aquellos grupos con las mayores frecuencias relativas acumuladas. Para ello después de determinar el grupo relevante individual se considerarán también como relevantes el grupo o grupos cuya diferencia en frecuencia relativa acumulada sea menor a un determinado umbral. A este grupo relevante (o grupos relevantes) lo denominaremos **grupo relevante acumulado**.

Después de establecer la selección del grupo relevante, y en consecuencia el o los grupos no relevantes, se procede a expandir la petición.

3.3 Expandir la petición

El proceso de expandir la petición tiene como objetivo reformular la petición inicial considerando la información que se obtenga de la petición inicial y de la que resulte del análisis a los pasajes relevantes y no relevantes previamente seleccionados. Esta nueva petición retroalimentará al módulo de recuperación de pasajes, para nuevamente generar otro conjunto de pasajes.

También es importante mencionar que para el proceso de expansión, el sistema de recuperación de pasajes que se utilice es visto como una caja negra, de tal forma que dicho proceso pueda aplicarse a cualquier sistema de recuperación de pasajes sin importar su arquitectura interna.

Como ya se explicó en la sección 2.3.2, la técnica de retroalimentación de relevancia se ha utilizado dentro de RI para modificar la petición original, ya sea agregando o eliminando términos, resultando esto del análisis realizado a los primeros documentos recuperados por el sistema de RI. También se mencionó que una de las técnicas más representativas dentro de la retroalimentación de relevancia es la técnica de Rocchio (sección 2.3.3). Sin embargo, debido a que la expresión (2.1.3) que denota a la técnica de Rocchio se basa en un modelo vectorial para fines del método propuesto, dicha técnica se adaptó para no depender del modelo basado en vectores. Esta adaptación de la técnica de Rocchio se detalla enseguida.

Sea *Q* la pregunta inicial formada por un conjunto de términos definida como:

$$Q = \{q_1, ..., q_k\} \tag{3.1.2}$$

A su vez P^Q es el conjunto de pasajes recuperados para *Q* definido como:

$$P^Q = \{P_1, P_2, ..., P_N\} \tag{3.1.3}$$

V^Q es el vocabulario obtenido de P^Q formado por un conjunto de términos representado como:

$$V^Q = \{t_1, t_2, ..., t_{|V^Q|}\} \tag{3.1.4}$$

p_k es el vector del *k-ésimo* pasaje recuperado denotado como:

$$p_k = \left\langle w_{t_1}, ..., w_{t_{|V^Q|}} \right\rangle \tag{3.1.5}$$

donde w_{t_i} es el peso del término t_i en el pasaje *k*.

Además R^Q es el conjunto de pasajes relevantes para *Q* y N^Q el conjunto de pasajes no relevantes para *Q*. Estos conjuntos cumplen con las siguientes expresiones:

$$R^Q \cap N^Q = \phi \tag{3.1.6}$$

$$P^Q = R^Q \cup N^Q \tag{3.1.7}$$

De este conjunto de pasajes relevantes y no relevantes se construye una matriz booleana para el grupo de pasajes relevantes y otra para el grupo de pasajes no relevantes. Dicha matriz está formada por el vocabulario y los pasajes de cada grupo. Cada intersección de la matriz asocia la existencia o no de un término del vocabulario en el pasaje.

Para cada término en la matriz se establece una suma promedio de las ocurrencias de éste en los pasajes, con respecto al total de pasajes por cada grupo, de esta manera se forma el vector promedio de términos para el grupo relevante (P^{R^Q}) y el no relevante (P^{N^Q}). Ambos vectores indican la ocurrencia promedio de los términos del vocabulario en cada grupo. Dichos vectores se expresan de la siguiente manera:

$$P^{R^Q} = \frac{1}{\left|R^Q\right|} \sum_{\forall p_i \in R^Q} p_i \qquad (3.1.8)$$

$$P^{N^Q} = \frac{1}{\left|N^Q\right|} \sum_{\forall p_i \in N^Q} p_i \qquad (3.1.9)$$

Posteriormente se realiza la operación entre los vectores anteriores tomando en cuenta los coeficientes de Rocchio: β y γ, los cuales multiplican a P^{R^Q} y P^{N^Q}, respectivamente, para formar el vector que indica los valores de cambio de los términos del vocabulario denotado por:

$$V^{C^Q} = \beta P^{R^Q} - \gamma P^{N^Q} \qquad (3.1.10)$$

Donde: $$\beta + \gamma = 1 \qquad (3.1.11)$$

y $$V^{C^Q} = \left\langle v_{t_1}, ..., v_{t_{|V^Q|}} \right\rangle \qquad (3.1.12)$$

donde v_{t_i} es el valor del término t_i en el elemento *i* de V^{C^Q}.

En el proceso normal de la técnica de Rocchio se realiza una suma de vectores entre la petición inicial y el V^{C^Q} para construir el nuevo vector de la petición como lo indica la expresión (2.1.3). En esta tesis la técnica de Rocchio se adapta, es decir, con los valores de V^{C^Q} se obtienen dos conjuntos T^R y T^N.

T^R es un conjunto de términos del vocabulario V^Q que tienen el valor máximo entre los elementos de V^{C^Q} y son los términos para agregarse a la petición inicial. T^R se denota como:

$$T^R = \left\{ t_i \in V^Q \mid t_i \notin Q \wedge v_{t_i} = \max(V^{C^Q}) \right\} \quad (3.1.13)$$

T^N es un conjunto de términos del vocabulario V^Q que tienen el valor mínimo entre los elementos de V^{C^Q} y son los términos a eliminar de la petición inicial. T^N se denota como:

$$T^N = \left\{ t_i \in V^Q \mid t_i \in Q \wedge v_{t_i} = \min(V^{C^Q}) \right\} \quad (3.1.14)$$

Se consideran como conjuntos a T^R y T^N porque pueden existir empates entre los valores máximos y mínimos, y esto provoque que sea más de un término al que le correspondan dichos valores.

Finalmente, la nueva petición se forma de la unión de los términos en T^R, y eliminando los términos T^N de la petición inicial Q, formando la nueva petición Q^1 se denota como:

$$Q^{'1} = (Q \cup T^R) - T^N \quad (3.1.15)$$

El método propuesto puede aplicarse iterativamente, ya que los pasajes se pueden volver a recuperar con la nueva petición y repetir el proceso. En la siguiente sección se detallan los experimentos y evaluaciones realizadas al método propuesto.

Capítulo 4

Experimentos y Resultados

Dado que el método propuesto consta de dos formas para realizar el proceso de la selección de pasajes, nuestros experimentos se dividen en dos evaluaciones. Primero evaluamos la forma de seleccionar los pasajes con el **método tradicional** al determinar con un umbral los pasajes relevantes. Posteriormente, evaluamos el **método basado en agrupamiento** donde la selección de pasajes se realiza con base en una técnica de agrupamiento.

4.1 Corpus

Los experimentos fueron desarrollados considerando el corpus en Español del CLEF[2] (Cross Language Evaluation Forum). Éste es un foro donde se evalúan tareas de Recuperación de Información, Extracción de Información y Búsqueda de Respuestas. El corpus de documentos que se consideró para los experimentos consiste de 454,045 noticias de la agencia EFE de los años 1994 y 1995.

El conjunto de preguntas que se utilizó es de 2005 y se compone de 123 preguntas factuales, para las cuales existe su respuesta. Dicho conjunto de preguntas se dividen en seis grupos según el tipo de respuesta esperada:

[2] http://www.clef-campaign.org

21 de cantidad, 16 de fecha, 25 de lugar, 15 de misceláneo, 20 de organización y 26 de persona.

Como parte de la evaluación fue necesario contar con las respuestas correctas de las preguntas. Éstas fueron tomadas de los corpora generados por los organizadores del CLEF.

4.2 Métricas de evaluación

Las métricas usadas en la evaluación son la cobertura, la redundancia y el ruido. Estas métricas se basan en las nociones de las expresiones (2.1.1) y (2.1.2). Sin embargo, para facilitar su explicación definiremos el concepto de preguntas factibles Q^F, es decir, el conjunto de aquellas preguntas que tienen al menos un pasaje con la respuesta correcta.

$$Q^F = \left\{ q \in Q \middle| R^S_{P,q,n} \cap A_{P,q} \neq \phi \right\} \qquad (4.1.1)$$

Donde Q es el conjunto de preguntas, P la colección de pasajes, $A_{P,q}$ el subconjunto de P el cual contiene respuestas correctas para $q \in Q$, y $R^S_{P,\,q,\,n}$ representa los primeros n pasajes en P recuperados por un sistema de recuperación S dada una pregunta q.

La cobertura (*cobertura*^{S}*(Q,P,n)*) tiene por objetivo proporcionar el porcentaje de preguntas factibles dentro de la colección, se denota por la expresión (4.1.2).

$$cobertura^S(Q,P,n) = \frac{\left|Q^F\right|}{\left|Q\right|} \qquad (4.1.2)$$

La redundancia (*redundancia*^{S}*(Q,P,n)*) tiene por objetivo proporcionar el número promedio de pasajes relevantes de las preguntas factibles.

$$redundancia^{S}(Q,P,n) = \frac{\sum_{\forall q \in Q} \left| R^{S}_{P,q,n} \cap A_{P,q} \right|}{\left| Q^{F} \right|} \qquad (4.1.3)$$

Una última medida propuesta para la evaluación de nuestros métodos fue una medida del ruido. En nuestro caso particular, el ruido es toda aquella información no relevante que será entregada al proceso de extracción de la respuesta. Mientras menor sea el ruido mayores posibilidades de extraer la respuesta correcta. Así el ruido (*ruido^{S}(Q,P,n)*) mide el porcentaje de pasajes no relevantes dentro de las preguntas factibles y se expresa en (4.1.4).

$$ruido^{S}(Q,P,n) = \frac{\sum_{\forall q \in Q^{F}} \left| R^{S}_{P,q,n} - A_{P,q} \right|}{\left| Q^{F} \right|} \qquad (4.1.4)$$

A través de un ejemplo se detallan estás métricas de la evaluación. La tabla 4.1 muestra un conjunto de pasajes recuperados para distintas preguntas, el número 1 en la matriz indica que es un pasaje relevante y el 0 que es no relevante. Cabe señalar que se recuperaron distinto número de pasajes para las preguntas 2 y 4, por lo que los cuadros con relleno gris en la tabla 4.1 indican que no se cuentan con dichos pasajes.

Tabla 4.1 Ejemplo para aplicar las métricas

	Pasaje1	Pasaje2	Pasaje3	Pasaje4	Pasaje5
Pregunta1	0	0	0	0	0
Pregunta2	1	0	0		
Pregunta3	1	1	0	1	0
Pregunta4	0	0	1	0	

Para el ejemplo de la tabla 4.1 tenemos los siguientes resultados:

$$cobertura = \frac{3}{4} = 0.75$$

$$redundancia = \frac{1+3+1}{3} = 1.6$$

$$ruido = \frac{2+2+3}{12} = 0.583$$

La cobertura indica que el 75% de las preguntas tienen al menos un pasaje relevante. En contraste, el 25% de las preguntas no tienen pasajes con la respuesta correcta.

La redundancia por otro lado indica que en promedio se tienen 1.6 pasajes relevantes para las preguntas factibles. Un aumento en la redundancia beneficia al proceso de la extracción de la respuesta en un sistema BR, ya que es más probable extraer la respuesta correcta.

El ruido indica que el 58.3% del total de pasajes recuperados para las preguntas factibles son no relevantes.

4.3 Método base (sin expansión)

Para tener un punto de referencia se realizó un experimento sin realizar el proceso de expansión. A este experimento lo denominaremos el método base. Para este método base se tomaron las 123 preguntas consideradas en el corpus, estas preguntas convertidas en peticiones (sin considerar el tipo de pregunta y las palabras vacías) son dadas al sistema de recuperación de pasajes para obtener 30 pasajes (cada pasaje formado por una frase) para cada una de las peticiones.

Como resultados para este método base se obtuvo una cobertura del 80%, una redundancia de 7.28 y un valor de ruido del 76%.

La cobertura nos indica que sólo el 20% de preguntas no son factibles. La redundancia indica que se tiene en promedio siete pasajes relevantes para las preguntas factibles. Mientras que el ruido refleja que un 76% de la información recuperada no genera información relevante. Un aumento en cobertura y redundancia, y la disminución del ruido, son los comportamientos ideales para mejorar estos resultados.

Estos datos marcan el punto de referencia para comparar los resultados alcanzados con el método propuesto aplicando la expansión de la petición.

La tabla 4.2 muestra los resultados del método base por grupos, esto para tener una referencia del comportamiento según el tipo de respuesta esperada, ya que en las siguientes secciones en algunos experimentos se mostrará el comportamiento por grupos después de la expansión.

Tabla 4.2 Método base por grupos

Grupo	**Método base**	
	cobertura	redundancia
Cantidad	0.67	3.5
Fecha	0.63	3.5
Lugar	0.96	8.42
Misceláneo	0.73	6.45
Organización	0.90	8.61
Persona	0.85	9.5

4.4 Método tradicional

La expansión de la petición aplicando al método tradicional se realizó a partir de un conjunto de 30 pasajes que se recuperaron para cada petición, se estableció un umbral n=10 para determinar los pasajes relevantes. De esta manera, los pasajes uno al 10 se seleccionó como relevantes. Para incrementar la diferencia entre los relevantes y los no relevantes se seleccionaron como no relevantes los últimos diez pasajes, es decir, los

pasajes 21 al 30. Con estos pasajes, se aplicó la técnica de Rocchio adaptada para lograr la extracción de términos para la expansión de la petición.

Para llevar a cabo el método tradicional fue necesario determinar los valores de los coeficientes de Rocchio. A continuación se detallan los experimentos realizados para determinarlos.

4.4.1 Estimación de los coeficientes de Rocchio

Estos coeficientes de Rocchio son importantes ya que el valor de β y γ indican cómo se aplica la técnica de Rocchio. β y γ como coeficientes de Rocchio son los que proporcionan un factor de peso a los términos de los pasajes relevantes y a los no relevantes, respectivamente, para posteriormente elegir los términos candidatos para la expansión. Los posibles valores para β y γ son entre 0 y 1. Si **β=0.5** y **γ=0.5**, se indica que basta con que aparezca un término de los relevantes en los no relevantes para no considerarse dicho término como relevante en el proceso de expansión. Mientras que si **β=0.75** y **γ=0.25**, se indica que un término que aparece una vez en los relevantes, basta con que aparezca tres veces en los no relevantes para no considerarse dicho término como relevante en el proceso de expansión.

La tabla 4.3 muestra un resumen de los resultados que se obtuvieron en cuanto a las medidas de *cobertura* y *redundancia* a 30 pasajes. La evaluación para este experimento se realizó por grupos y de forma global, en la mayoría de estos grupos se lograron mejores resultados con **β=0.75** y **γ=0.25** en cuanto a *cobertura* y *redundancia*.

Tabla 4.3 Cobertura y redundancia para β y γ en el método tradicional

Grupo	β=0.5 y γ=0.5		β=0.75 y γ=0.25	
	cobertura	redundancia	cobertura	redundancia
Cantidad	**0.48**	2.50	0.38	**2.75**
Fecha	0.44	**8.43**	**0.63**	6.00
Lugar	0.76	10.2	**0.93**	**12.3**
Misceláneo	0.47	9.71	**0.53**	**13.4**
Organización	0.75	13.2	**0.80**	**13.9**
Persona	0.69	9.11	**0.77**	**12.6**
Global	0.62	9.30	**0.70**	**11.1**

4.4.2 Resultados

Dadas las características del método de expansión propuesto se observó el comportamiento de la expansión al aplicarlo iterativamente. La tabla 4.4 muestra los resultados de los experimentos realizados para la expansión de peticiones al aplicarla en dos iteraciones.

Tabla 4.4 Resultados del método tradicional

Tipo de corrida	A 30 pasajes		
	cobertura	redundancia	ruido
Método base	***0.80***	*7.28*	*0.76*
Método tradicional Iteración 1	0.70	**11.1**	**0.63**
Método tradicional Iteración 2	0.72	**11.1**	**0.63**

Como puede observarse en la tabla 4.4, los resultados para la cobertura son mejores en la iteración 2 que la iteración 1 al aplicar el método tradicional, obteniendo un 72% de cobertura a los 30 pasajes. También se observa que en comparación con el método base se pierde un 10% de cobertura, lo que significa que se pierden algunas preguntas que antes eran contestadas en algún pasaje. Esto puede deberse a que la información era muy escasa, por lo que al realizar la expansión se desvió la información provocando una ligera pérdida de preguntas factibles.

Los resultados en la redundancia incrementan en un promedio de siete a 11 veces el número de pasajes con la respuesta correcta a la pregunta.

Los resultados para el ruido indican una reducción del 17% del método tradicional con respecto al método base, lo que implica que se disminuyó el número de pasajes no relevantes para las preguntas factibles.

La redundancia y el ruido muestran el efecto que provoca la expansión de la petición, es decir, hubo un aumento en pasajes relevantes y por consiguiente disminución en el número de pasajes no relevantes, para las preguntas factibles. En la cobertura se perdieron algunas preguntas que eran factibles con respecto al método base, muy probablemente hayan sido preguntas que antes de la expansión tenía escasos pasajes relevantes, y la información extraída de estos no fue suficiente para que la expansión reformulara la petición adecuadamente.

De los resultados de las iteraciones del método tradicional si hubo ciertas mejoras con respecto a los del método base, pero obsérvese que resulta muy arbitrario considerar –en base a un umbral– los pasajes relevantes y los no relevantes, ya que no se garantiza la relevancia de dichos pasajes según su posición en las lista de los pasajes recuperados.

Para esto en la siguiente sección se detallan los experimentos y resultados obtenidos al aplicar la segunda forma propuesta de seleccionar los pasajes, es decir, mediante la técnica de agrupamiento.

4.5 Método basado en Agrupamiento

Este método consiste en aplicar una técnica de agrupamiento a los pasajes recuperados. A partir de los grupos de pasajes calculados se elige un grupo relevante, con el cual se aplicará el proceso de expansión de la petición. Para aplicar este método también se calcularon los coeficientes de Rocchio de manera experimental. A continuación se detalla estos experimentos.

4.5.1 Estimación de los coeficientes de Rocchio

La tabla 4.5 muestra un resumen de los resultados que se obtuvieron en cuanto a las medidas de *cobertura* y *redundancia*, a 30 pasajes. Se evaluaron por grupos de preguntas y de forma global, los mejores resultados en cuanto a cobertura y redundancia se dan en el caso **β=0.75** y **γ=0.25**.

Tabla 4.5 Cobertura y redundancia para β y γ con agrupamiento

Grupo	**β=0.5** y **γ=0.5**		**β=0.75** y **γ=0.25**	
	cobertura	redundancia	cobertura	redundancia
Cantidad	0.52	2.55	**0.52**	**3.09**
Fecha	0.69	4.18	**0.69**	**4.18**
Lugar	**0.80**	**13.8**	0.76	13.0
Misceláneo	0.73	11.5	**0.73**	**11.5**
Organización	0.85	10.0	**0.90**	**10.7**
Persona	0.88	12.3	**0.85**	**14.2**
Global	0.75	9.79	**0.76**	**10.6**

4.5.2 Evaluación

En esta sección se muestran experimentos y resultados del método basado en agrupamiento. Para el agrupamiento se uso la técnica de agrupamiento **K-Means**.

Para aplicar el algoritmo K-Means es necesario decidir *a priori* el número de grupos deseado. En nuestro caso se eligió una *k*=3. Esta decisión se tomó con base en la suposición de que es posible ubicar en un solo grupo los pasajes relevantes pero no así los no relevantes. Es decir, es posible reunir en un grupo la mayoría de los pasajes relevantes, puesto que son los pasajes similares, posiblemente con la respuesta. Pero asumir que en un sólo grupo quedarán los pasajes no relevantes es demasiado, quizá no sea correcto, ya que es de esperarse que dichos pasajes contiene información muy variada, entonces se optó considerar 3 grupos.

En la sección 3.2.2 se indicaron los criterios para la selección del grupo relevante. Respecto al segundo criterio se usó un umbral de similitud del 5% del total de términos de la petición. Como se recordará este umbral permite considerar más de un grupo para confirmar el conjunto de pasajes relevantes.

De la misma manera que el experimento del método tradicional se realizaron dos iteraciones de la expansión de la petición, aplicando la técnica de agrupamiento a los pasajes recuperados.

La tabla 4.6 muestra las evaluaciones de las iteraciones especificando el criterio aplicado en la selección del grupo relevante, ya sea individual o acumulado, después de la técnica de agrupamiento.

Entre las iteraciones 1 y 2, se observa una mejora en la mayoría de los valores para los pasajes recuperados en la iteración 1 con el criterio de grupo relevante acumulado, teniendo una cobertura del 76%. Manteniendo una redundancia de un promedio de 10 veces de que se halle la respuesta, y reduciendo el ruido en un 14% con respecto al método base, indicando que disminuyó en ese porcentaje el número de pasajes no relevantes.

Tabla 4.6 Resultados del método basado en agrupamiento

Tipo de corrida	**Grupo relevante individual**			**Grupo relevante acumulado**		
	cobertura	redundancia	ruido	cobertura	redundancia	ruido
Método base	*0.80*	*7.28*	*0.76*	***0.80***	*7.28*	*0.76*
Método basado en agrupamiento iteración 1	0.73	8.79	0.71	0.76	**10.6**	**0.65**
Método basado en agrupamiento iteración 2	0.69	9.66	0.69	0.76	10.3	0.66

Estos resultados con respecto al método base mantienen la tendencia de ganar en redundancia, disminuir el ruido y perder en cobertura. Sin embargo,

las dos primeras tendencias son importantes mejoras, sobre todo la del ruido ya que aumenta las probabilidades de extraer la respuesta correcta.

También, con el uso de los criterios de selección del grupo relevante y no relevante después del agrupamiento, se destaca que se dan mejores resultados al permitir la unión de otros grupos que hayan resultado del agrupamiento, que considerar a un grupo como relevante.

En general los resultados de este método basado en agrupamiento con respecto al método base se pueden considerar como mejoras al proceso de recuperación de pasajes, con la aplicación de la expansión de la petición.

4.6 Comparación de los métodos

La diferencia entre el método tradicional y el basado en agrupamiento consiste en la forma de seleccionar los pasajes relevantes y los no relevantes, para a partir de estos, realizar la expansión de la petición.

La tabla 4.7 resume los datos en cobertura, redundancia y ruido que se obtuvieron en ambos métodos a los 30 pasajes. Los resultados que se muestran del método basado en agrupamiento son los que corresponden al hacer uso del grupo relevante acumulado.

Se observa en dicha tabla que en ambos métodos con respecto al método base se disminuye en cobertura, lo que indica que se perdieron algunas preguntas que en el método base tenían pasajes relevantes.

En cuanto a redundancia con ambos métodos se supera al método base. Esto indica que se tienen más pasajes relevantes después de aplicar la expansión, es decir, ha aumentado el número de pasajes donde puede hallarse la respuesta correcta.

Con la medida del ruido, en ambos métodos hubo una disminución en el porcentaje con respecto al método base. Lo que indica que el proceso de expandir la petición ha logrado eliminar información no relevante de las preguntas que en un inicio son factibles.

Tabla 4.7 Resultados de los métodos

Tipo de corrida	A 30 pasajes		
	cobertura	redundancia	ruido
Método base	***0.80***	*7.28*	*0.76*
Método tradicional iteración 1	0.70	**11.1**	**0.63**
Método tradicional iteración 2	0.72	**11.1**	**0.63**
Método basado en agrupamiento iteración 1 (+)	0.76	10.6	0.65
Método basado en agrupamiento iteración 2 (+)	0.76	10.3	0.66

Los cambios mostrados entre una iteración y otra indican que una tercera iteración no es necesaria, ya que los resultados ya no parecen ser favorables desde la segunda iteración para ambos métodos.

Finalmente, como conclusión de los resultados mostrados, el comportamiento tanto del método tradicional como el basado en agrupamiento mejoran la calidad de los pasajes recuperados, aumentando el número de pasajes relevantes –y con ello la redundancia– y disminuyendo el ruido. Ahora bien, a pesar de la pérdida en cobertura, de un 80% (método base) a un 76% (método basado en agrupamiento) se tienen muchos más elementos para poder extraer un mayor número de respuestas correctas. La pérdida en cobertura se debe a que no existen elementos para reformular la petición apropiadamente.

Sin embargo, es necesario hacer más experimentos para comprobar si es posible adecuar nuestro método de expansión de la petición a esta situación.

Capítulo 5

Conclusiones y Trabajo Futuro

5.1 Conclusiones

El presente trabajo analizó la aplicación del proceso de la expansión de la petición orientado a las necesidades de la búsqueda de respuestas.

Para ello se experimentó con la idea de seleccionar pasajes apropiados para realizar la expansión. Dos métodos fueron propuestos: uno al aplicar la idea tradicional de seleccionar los primeros pasajes recuperados, y otro al aplicar una técnica de agrupamiento. En ambos métodos de selección se usó la retroalimentación ciega de relevancia con una adaptación de la técnica de Rocchio para realizar la expansión de la petición.

Los métodos propuestos no involucran ningún tipo de recurso lingüístico para el análisis de los pasajes o de los términos candidatos para la expansión, sólo se limita a las ocurrencias de los términos en los pasajes relevantes y en los no relevantes, además de su relación con los términos de la petición original. Lo anterior es una ventaja del método ya que puede aplicarse a otros idiomas.

Los métodos propuestos demostraron un buen desempeño en las evaluaciones realizadas. Tanto en la redundancia como en el ruido se lograron mejoras interesantes. De esta manera con más pasajes conteniendo

la respuesta y con menos pasajes sin la respuesta se entrega al módulo de extracción de la respuesta información más apropiada. Sin embargo, en ambos métodos la cobertura disminuye, es decir, se perdieron las respuestas de algunas preguntas que se tenían antes de aplicar la expansión de la petición. Cabe resaltar que aquellas preguntas que se perdieron se debe principalmente a la falta de pasajes relevantes; bajo estas circunstancias es muy probable que el módulo de extracción de la respuesta tampoco tenga los elementos suficientes para determinar la respuesta correcta.

En resumen, se puede decir que el método propuesto generó resultados favorables lo que permite concluir que es un método que contribuirá a la recuperación de pasajes dentro de un sistema de BR.

5.2 Trabajo futuro

Como trabajo futuro del método propuesto sería válido aplicar otras medidas de similitud y técnicas de agrupamiento para la selección de pasajes, sobre todo para observar el cambio en la cobertura.

Además, se pudiera considerar en el proceso de agrupamiento el tipo de respuesta esperada, para especializar el agrupamiento de los pasajes con base a esta información.

Otro punto, a realizar en un futuro inmediato, es evaluar el método de expansión en un sistema de búsqueda de respuestas para comprobar el impacto que tendría sobre los resultados globales del sistema.

A. Tabla de Símbolos

Símbolo	Nombre	Expresa:
$\|\ \|$	Cardinalidad de un conjunto	Número o cantidad de los elementos constitutivos de un conjunto
$\in$	Pertenencia de conjuntos	$a \in S$ significa: a es elemento del conjunto S
$\notin$	Pertenencia de conjuntos	$a \notin S$ significa: a no es elemento del conjunto S
$\cup$	Unión de conjuntos	$A \cup B$ significa: el conjunto que contiene todos los elementos de A y también todos aquellos de B, pero ningún otro.
$\cap$	Intersección de conjuntos	$A \cap B$ significa: el conjunto que contiene todos aquellos elementos que A y B tienen en común.
$\forall$	Cuantificación universal	$\forall\ x$: $P(x)$ significa: $P(x)$ es verdadera para cualquier x
$\neq$	Diferente	$a \neq b$ significa: a es diferente de b
ϕ	Conjunto vacío	Conjunto que no tiene elementos
Σ	Sumatoria	$\sum_{k=1}^{n} a_k$ significa: $a_1 + a_2 + ... + a_n$ suma sobre ... desde ... hasta ... de
$\vee$	Disyunción lógica	O
$\wedge$	Conjunción lógica	Y
$\langle\ \rangle$	Delimitadores del vector	El vector $V=\langle v_1, v_2, \ldots v_n \rangle$ .significa: el vector V consiste de $v_1, v_2, \ldots v_n$
max()	Función máximo	$b = max(A)$ significa: b es el valor máximo de A, donde A puede ser un conjunto o un vector.
min()	Función mínimo	$b = min(A)$ significa: b es el valor mínimo de A, donde A puede ser un conjunto o un vector.

Bibliografía

Billerbeck B., Zobel J. (2004) Questioning Query Expansion: An Examination of Behaviour and Parameters. In *Proceedings of the Fifteenth Australasian Database Conference (ADC):* 69-76*,* Dunedin, New Zealand.

Bilotti M., Katz B., Lin J. (2004) What Works Better for Question Answering: Stemming or Morphological Query Expansion? In *Proceedings of the Special Interest Group on Information Retrieval (SIGIR) Workshop on Information Retrieval for Question Answering (IR4QA):*1-7, Sheffield, England.

Callan J.-P. (1994) Passage-Level Evidence in Document Retrieval. In *Proceedings of the 17th Annual ACM SIGIR International Conference on Research and Development in Information Retrieval*: 302-310, Dublin, Ireland.

Chang C. H., Hsu C. C. (1998) Hypertext Information Retrieval for Short Queries. In *Proceedings of the IEEE Knowledge and Data Engineering Exchange Workshop:* 1-8, Taipei, Taiwan.

Cui H., Wen J.-R., Nie J.-Y., Ma W.-Y. (2002) Probabilistic Query Expansion Using Query Logs. In *Proceedings of the 11th World Wide Web Conference*: 325-332, Honolulu, Hawaii, USA.

Ferrández A. (2004) *Tecnologías del Texto y Habla*, capítulo de Sistemas de pregunta y respuesta: 11-14. Edicions Universitat de Barcelona.

Gelbukh A. (2000) Lazy Query Enrichment: A Simple Method of Indexing Large Specialized Document Bases. In *Proceedings of the 11th International Conference and Workshop on Database and Expert Systems Applications:* 526-535, Greenwich, England.

Greenwood M. (2004) Using Pertainyms to Improve Passage Retrieval for Questions Requesting Information About a Location. In *Proceedings of the Special Interest Group on Information Retrieval (SIGIR) Workshop on Information Retrieval for Question Answering (IR4QA):* 1-6, Sheffield, England.

Grossman D., Frieder O. (2004) *Information Retrieval: Algorithms and Heuristics*. Kluwer Academic Publishers.

Harabagiu S., Moldovan D. (2003) *The Oxford Handbook of Computational Linguistics*, chapter Question Answering: 560-582. Oxford University Press.

Harman D. (1992) Relevance Feedback Revisited. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieva*l: 1-10, Copenhagen, Denmark.

Hsu M.-H., Tsai M.-F., Chen H.-H. (2006) Query Expansion with ConceptNet and WordNet: An Intrinsic Comparison. In *Proceedings of the Third Asia Information Retrieval Symposium*: 16-18, Singapore.

Klink S. (2001) Query Reformulation with Collaborative Concept-based Expansion. In *First International Workshop on Web Document Analysis*: 19-22, Seattle, Washington, USA.

Llopis F., Ferrandez A., Vicedo J. (2002) Passage Selection to Improve Question Answering. In *Proceedings of the Conference On Computational Linguistics (COLING) Workshop on Multilingual Summarization and Question Answering:* 1-6, Morristown, NJ, USA.

Mandala R., Tokunaga T., Tanaka H. (1999) Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion. In *Proceedings of the 22nd Annual ACM SIGIR International Conference on Research and Development in Information Retrieval*: 15-19, Berkeley, CA, USA.

Monz C. (2003) From Document Retrieval to Question Answering. Tesis Doctoral. Institute for Logic, Language and Computation, University of Amsterdam, Holland.

Navigli R., Velardi P. (2003) An Analysis of Ontology-based Query Expansion Strategies. In *Proceedings of Workshop on Adaptive Text Extraction and Mining at the 14th European Conference on Machine Learning:* 1-8, Cavtat-Dubrovnik, Croatia.

Negri M. (2004) Sense-based Blind Relevance Feedback for Question Answering. In *Proceedings of the Special Interest Group on Information Retrieval (SIGIR) Workshop on Information Retrieval for Question Answering (IR4QA):* 1-8, Sheffield, England.

Noguera E., Llopis F. (2007) Applying Query Expansion techniques to Ad Hoc Monolingual tasks with the IR-n system. *Working Notes for the Cross Language Evaluation Forum (CLEF) Workshop,* Budapest, Hungary.

Roberts I., Gaizauskas R. (2004) Evaluating Passage Retrieval Approaches for Question Answering. In *Proceedings of the 26th European Conference on Information Retrieval (ECIR)*: 72—84*,* Sunderland, U.K..

Sanchis E., Rosso P., Montes-y-Gomez M., Gomez J.-M., Buscaldi D., Villaseñor-Pineda L. (2007) JAVA Information Retrieval System: An n-gram Model-based System for Passage Retrieval. *Enviado al Journal de Information Retrieval.*

Tellex S., Katz B., Lin J.-J., Fernandez A., Marton G. (2003) Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering. In *Proceedings of the 26th Annual ACM SIGIR International Conference on Research and Development in Information Retrieval*: 41-47, Toronto, Canada.

Xu J., Croft B. W. (2000) Improving the Effectiveness of Information Retrieval with Local Context Analysis. In *ACM Transactions on Information Systems (TOIS)*: 79—112.