Esta tesis doctoral contiene un índice que enlaza a cada uno de los capítulos de la misma.
Existen asimismo botones de retorno al índice al principio y final de cada uno de los capítulos.

Ir directamente al índice

Para una correcta visualización del texto es necesaria la versión de Adobe Acrobat Reader 7.0 o posteriores

Aquesta tesi doctoral conté un índex que enllaça a cadascun dels capítols.
Existeixen així mateix botons de retorn a l'index al principi i final de cadascun deis capítols .

Anar directament a l'index

Per a una correcta visualització del text és necessària la versió d' Adobe Acrobat Reader 7.0 o posteriors.

IR-n: Un sistema de Recuperación de Información basado en Pasajes

Tesis Doctoral

Departamento de Lenguajes y Sistemas Informáticos

Universitat d'Alacant
Universidad de Alicante

Autor
Fernando Llopis Pascual

Directores
Dr. José Luis Vicedo González
Dr. Antonio Ferrández Rodríguez

Alicante, 29 de marzo de 2003

A Luisa, Irene y Fernando

Agradecimientos

Quisiera expresar mi agradecimiento a los directores de esta tesis, José Luis Vicedo y Antonio Ferrández, no sólo por la ayuda que me han prestado en este trabajo, sino por lo mucho que he podido aprender de ellos.

A Jesús Peral y Rafael Muñoz por sus comentarios siempre acertados. A Fernando Martínez, mi tocayo de la Universidad de Jaén, por su colaboración en los trabajos con el sistema ZPrise. A Borja Navarro, Armando Suárez, Paloma Moreda, Luisa Quereda, Patricio Martínez y Maximiliano Saiz por ser desinteresados conejillos de indias en el debut interactivo del sistema IR-n. A Carol Peters por su amabilidad y ayuda en las Conferencias CLEF.

En general, quisiera hacer una mención muy especial a todos mis compañeros sin excepción alguna del Grupo de Procesamiento del Lenguaje y Sistemas de Información de la Universidad de Alicante, sin cuyos ánimos y colaboración me habría resultado muy difícil la finalización de este trabajo. Al Departamento de Lenguajes y Sistemas Informáticos por su apoyo en la realización de mi labor investigadora y docente.

A mi padres, sé que sin vuestro esfuerzo jamás podría haber estado aquí para iniciar este trabajo.

Sobre todo, quiero agradecer a Luisa, mi mujer, el apoyo incondicional y la comprensión que me ha ofrecido durante todo este tiempo de trabajo, viajes, congresos y fines de semana perdidos. Confío poder compensarla, aunque se que va a ser difícil.

Y a mis hijos, Irene (a ti se debe el nombre del sistema realizado en esta tesis) y Fernando por vuestras ganas de jugar y por contagiarme con vuestra alegría.

Alicante, Marzo 2003 *Fernando Llopis Pascual*

Índice General

Índice de Tablas

Índice de Figuras

Lista de abreviaturas

Abreviatura	Significado
AvgP	Media de precisión no interpolada
BR	Búsqueda de Respuestas
CLEF	Cross Lingual Evaluation Forum
EFE	Colección de documentos agencia EFE
EI	Extracción de Información
MUC	Message Understanding Conference
LATimes	Colección de documentos Los Angeles Times
PLN	Procesamiento del Lenguaje Natural
RI	Recuperación de Información
RID	Recuperación de Información basada en el análisis del documento completo
RP	Recuperación de Pasajes
SID	Selección Interactiva de Documentos
TREC	Text REtrieval Conference

1. Introducción

"El hombre es la única especie del planeta que ha inventado una memoria comunal que no está almacenada ni en nuestros genes ni en nuestros cerebros, el almacén de esta memoria se llama biblioteca" (Sagan, 1982).

Tradicionalmente, las bibliotecas han sido el repositorio donde se almacena la sabiduría acumulada por la humanidad durante las diferentes épocas. Alejandría, ciudad egipcia a orillas del Mediterráneo fundada por Alejandro Magno en los primeros años del siglo III a.c. fue célebre, además de por su gran faro, por disponer de la más completa e importante biblioteca del mundo conocido. No se conoce con seguridad el tamaño de la biblioteca, pero se estima que llegó a disponer de unos 750.000 volúmenes, entre códices, manuscritos, tablillas de arcilla y papiros. Esta cantidad puede parecer insignificante frente a los 23 millones de volúmenes que dispone la Biblioteca del Congreso de los Estados Unidos. No obstante, esos 750.000 volúmenes representaban prácticamente la totalidad del conocimiento humanístico y científico de su época.

La forma en la que los responsables de dicha biblioteca consiguieron tal cantidad de documentos fue variada. Por una parte se enviaban agentes al exterior con el objetivo de adquirir bibliotecas enteras. Por otra parte, se tiene constancia que la colección de la biblioteca creció espectacularmente gracias a una estrategia de piratería intelectual que hoy en día escandalizaría a las editoriales. Cada barco que pasaba por Alejandría, uno de los más importantes puertos de la antigüedad, era abordado y se incautaba cualquier manuscrito que transportara. Los manuscritos eran copiados y posteriormente devueltos a sus propietarios.

Estos hechos presentan la relativa facilidad con la que se pueden recopilar grandes cantidades de documentos. No obstante, para que tal cantidad de documentos sea realmente útil, se deben definir una serie de mecanismos que faciliten la localización de la información que se necesita en cada momento.

Para ello, durante más de 4.000 años los hombres han desarrollado y utilizado diversas técnicas para organizar y clasificar la información de la que disponían. Por ejemplo, dentro de las funciones de los bibliotecarios de la biblioteca de Alejandría, se hallaba, además de ser custodios de los volúmenes allí depositados, la organización de dichos volúmenes.

Ejemplos de esta organización son los archivadores y fichas clasificadas por materias y/o autores que se utilizan en las bibliotecas para facilitar la localización de algún determinado libro. También cabe citar, la propia ubicación de los libros en la biblioteca, como forma de facilitar el hallazgo de libros de alguna temática en concreto. Estos métodos de organización se basan fundamentalmente en el trabajo manual, tanto el de construcción de los elementos que facilitan la búsqueda, como el propio proceso de búsqueda que debe realizar el usuario para localizar la información deseada.

Este modelo de organización es, además de costoso, en ocasiones incompleto e ineficaz. Acudir a una biblioteca conociendo el nombre del autor o el título del libro deseado es sinónimo de éxito en la búsqueda, siempre que la biblioteca disponga de dicho libro. El mecanismo de esta búsqueda consiste en acceder bien al archivo de fichas ordenadas por autor o por título y localizar la ficha que suministra la ubicación del libro buscado. Este método de búsqueda es eficaz cuando se conoce exactamente uno de estos datos. No obstante si se conocía de forma parcial el título o se buscan libros de una temática en concreto, se debe realizar una búsqueda secuencial en el archivador organizado por materias, o bien acudir a la zona de la biblioteca donde se hallan albergados los libros que tratan sobre dicha temática.

Los primeros trabajos de informatización de las bibliotecas, solucionaron de forma parcial estos problemas. El hecho de almacenar en una base de datos la misma información que se escribía en las fichas, incrementaba notablemente las posibilida-

des de búsqueda. En primer lugar facilitaba el hecho en sí de la búsqueda, cambiando el acceso manual por la interacción con el sistema a través de un ordenador. Por otro lado, ya no era necesario conocer el título exacto del libro o el nombre del autor, sino únicamente algunas palabras del mismo para poder acceder a todos los libros que contuviesen en su título o nombre de autor dichas palabras.

A su vez, esto facilitaba el hecho de buscar libros que trataran sobre un tema, por ejemplo "Juan de Austria", "Visual Basic", etc. No obstante, obligaba a que estas palabras estuviesen contenidas en el título, o por lo menos en una serie de palabras clave que se asociaban a cada libro. Así, una búsqueda por palabras como "Felipe II", proporcionaría con total seguridad una serie de libros, ya que en España existen una gran cantidad de libros publicados sobre dicho tema y muchos de ellos contienen dichas palabras en el título. Pero, probablemente una búsqueda por personajes menos conocidos tales como "General Custer" o "William Wallace" podría ser infructuosa aun disponiendo de libros en la biblioteca que traten dicho tema.

Este problema podría ser solucionado si los sistemas de búsqueda permitieran localizar libros o documentos no a través de su título o nombre del autor, sino por su contenido completo. Es decir, poder localizar libros que traten determinado tema, sin necesidad que dicho tema esté reflejado en el título o incluso en los conceptos clave asociados al libro. Para ello es necesario que los documentos se hallen en formato electrónico. Ejemplos de repositorios de documentos en formato electrónico son las bibliotecas digitales y sobre todo el fenómeno Internet.

Una biblioteca digital es una colección organizada de documentos en formato digital. Este formato abre un amplio espectro de posibilidades. En primer lugar, incrementa las posibilidades de localización de documentos. En segundo lugar, permite que el usuario acceda a esta información sin tener que desplazarse físicamente a la biblioteca. Y en último lugar, disminuye la cantidad de espacio necesario para albergar una biblioteca. El almacenamiento de los documentos en forma digital supone, por ejemplo, que la información contenida en los libros de una sala de una bibliote-

ca, que ocupan un número de metros cuadrados considerable, se pueda almacenar en una serie de pequeños discos magnéticos.

Por otra parte, mientras que toda biblioteca, sea digital o no, se considera dotada de cierta estructura, Internet es una gran colección desorganizada de documentos en formato electrónico. Como mayor ventaja que ha supuesto Internet, cabe citar el imparable crecimiento de la disponibilidad de información para cualquier usuario. Como mayor inconveniente destaca su total falta de estructura, lo que ha convertido los métodos tradicionales de búsqueda, por autor o título, en poco operativos.

El éxito de Internet como fuente de información, las buenas perspectivas de las bibliotecas digitales, así como el hecho de que se ha generado una sociedad que se alimenta de la información, ha provocado el relanzamiento de la investigación en técnicas que permitan, utilizando formatos electrónicos, almacenar información y acceder a la misma de forma automática.

1.1 Obtención de información de forma automática

El objetivo básico de todas las técnicas de obtención o búsqueda de información de forma automática es el de suministrar, de forma eficaz y eficiente, aquella información solicitada por los usuarios. Existen diferentes aproximaciones o técnicas de búsqueda automática de información. Las diferencias entre estas técnicas radican principalmente en el enfoque que se da a este tratamiento automático y, sobre todo, en el objetivo final que se desea conseguir. Dicho objetivo viene siempre fijado por las necesidades de información que puede tener un usuario en un momento determinado. Estos objetivos diferencian las tres técnicas más conocidas de búsqueda de información de forma automática: la *Recuperación de Información* (o *Information Retrieval*, en adelante RI), la *Extracción de Información* (o *Information Extraction*, en adelante EI) y la *Búsqueda de Respuestas* (o *Question Answering*, en adelante BR). En los siguientes apartados se realizará una introducción a estas tres técnicas.

1.1.1 La Recuperación de Información (RI)

La RI es la tarea de localizar, dentro de una colección o corpus, aquellos documentos que son relevantes a una necesidad de información de un usuario (Smeaton, 1997). A pesar de que en esta definición no se concreta, la RI es una técnica que hace referencia a la recuperación automática de información. Una buena definición del término se encuentra en Lancaster (1968): "Un sistema de Recuperación de Información no informa a (no cambia el conocimiento de) un usuario con respecto a una pregunta que éste realiza, sino que meramente informa de la existencia (o no existencia) de documentos relacionados con dicha pregunta". En (Salton, 1989) ya no se limitan los objetivos de un sistema de RI a determinar la relevancia o no de un documento con respecto a una pregunta, sino que se añade el concepto de orden (*ranking*), de forma que se otorga una puntuación a cada documento en función de su similitud o relevancia con respecto a la pregunta. Esta puntuación permite ordenar de mayor a menor relevancia los documentos de una colección con respecto a dicha pregunta. Este modelo se puede ver en la figura 1.1, en la que aparecen los conceptos fundamentales del proceso: la pregunta, la colección de documentos disponibles y el sistema de RI que permite seleccionar aquellos que son relevantes. Una definición más genérica sobre RI se puede encontrar en Baeza-Yates y Ribeiro-Neto (1999), donde la RI se define como "una técnica relacionada con la representación, almacenamiento, y acceso a campos de información".

Un ejemplo de pregunta que procesaría un sistema de RI sería la siguiente: "Encuentra documentos que contengan información relevante sobre la arquitectura desarrollada en Berlín con posterioridad a la caída del muro", o bien de forma más simple, "Arquitectura en Berlín".

Puede que actualmente el concepto de RI como tal no sea lo suficientemente conocido por todos los usuarios de este tipo de sistemas, no obstante, éstos sí que conocen uno de los más claros ejemplos de sistemas de RI existentes: los motores de búsqueda de documentos en Internet. Para estos sistemas, la colección donde se deben localizar los documentos relevantes es el conjunto

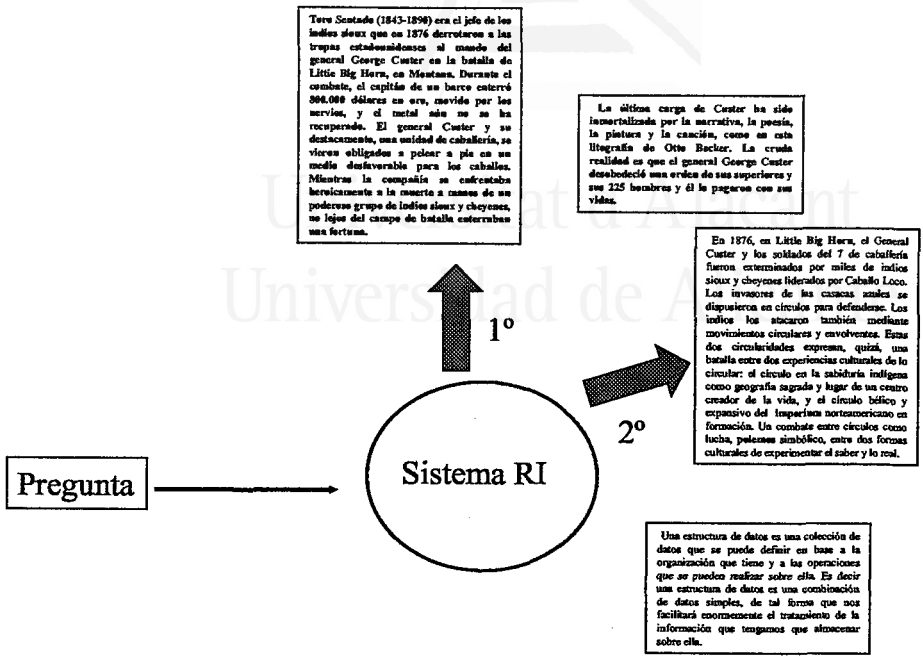


Figura 1.1. Sistema de recuperación de información

de páginas HTML y documentos disponibles en Internet. Pueden servir como ejemplo, los buscadores más utilizados actualmente como Google[1] y Yahoo[2].

Uno de los principales foros de investigación en sistemas de RI lo constituyen las series anuales de conferencias Text RETrieval Conference (TREC[3]) y más recientemente las Cross Lingual Evaluation Forum (CLEF[4]). En estas conferencias se diseñan una serie de tareas que tienen como principal objetivo evaluar y comparar el rendimiento de diversos sistemas de RI. Las actas de ambas conferencias permiten comprobar la evolución de las investigaciones en RI.

[1] http://www.google.com
[2] http://www.yahoo.com
[3] http://trec.nist.gov
[4] http://www.clef-campaign.org

1.1.2 La Extracción de Información (EI)

La EI es una técnica relacionada con la RI, pero mientras ésta recupera documentos, la EI procesa textos y los convierte de forma parcial o total en un conjunto de datos relevantes (Smeaton, 1997). Así, la diferencia entre sistemas de RI y EI radica en que los primeros localizan documentos relevantes mientras que los segundos localizan y extraen información relevante. En Wilks (1997) se incorpora el concepto de información pre-especificada, es decir, el usuario no sólo solicita el tema que él considera relevante sino los items asociados a ese tema que realmente le interesan. Con ello, al concepto de búsqueda de información, se añade el concepto de plantilla que debe rellenar de forma automática el sistema (ver figura 1.2).

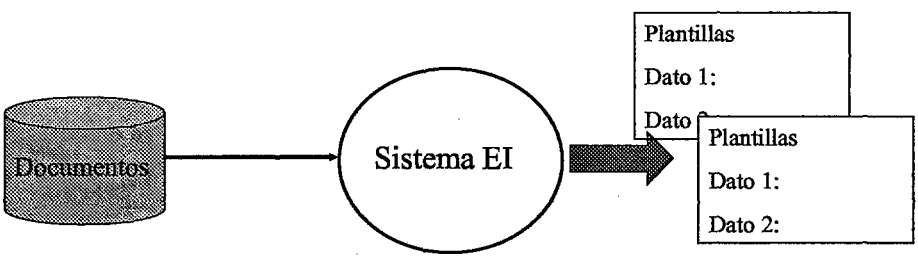


Figura 1.2. Sistema de extracción de información

Así, los sistemas de EI tratan de solucionar el problema principal de la RI. Una vez que el usuario ha recibido la lista de documentos relevantes todavía le queda pendiente una ardua tarea por realizar. Esta tarea consiste en comprobar primero si esos documentos contienen la información que busca y, a continuación, intentar localizar el lugar dentro del mismo en el que se halla dicha información.

La idea original de la EI era la de analizar textos de dominios no restringidos, pero actualmente los sistemas de EI trabajan con

textos en dominios específicos, ya que en textos no restringidos su eficacia suele ser limitada. Además, los sistemas de EI no permiten el tratamiento de preguntas arbitrarias, sino que el tipo de información requerida necesita ser definida de forma previa al desarrollo del sistema.

Un ejemplo de sistema de EI serían los trabajos realizados con el sistema EXIT (Llopis et al., 1998). Este sistema tiene como objetivo la obtención de información referente a las operaciones de compra-venta que se reflejan en una serie de textos notariales. Esta información consiste en conocer los nombres de las personas o entidades que intervienen en la operación, así como su rol (comprador, vendedor, representante, etc.), el bien y el precio de venta del mismo.

Al igual que la investigación en RI ha sido estudiada en las conferencias TREC y CLEF, la serie de conferencias Message Understanding Conference (MUC) ha constituido el forum principal de promoción, comparación y evaluación de la tecnología desarrollada en los sistemas de EI.

1.1.3 La Búsqueda de Respuestas (BR)

El hecho que los sistemas de EI tengan una excesiva dependencia del dominio, y principalmente, un creciente interés en sistemas que afrontaran con éxito la tarea de localizar respuestas concretas a preguntas arbitrarias formuladas por los usuarios en grandes volúmenes de información, han dejado la puerta abierta a la aparición de un nuevo campo de investigación conocido como BR.

La BR se puede definir como aquella tarea automática realizada por ordenadores que tiene como finalidad la de encontrar respuestas concretas a necesidades precisas de información formuladas por los usuarios (ver figura 1.3). Los sistemas de BR son especialmente útiles en situaciones en las que el usuario final necesita conocer un dato muy específico y no dispone de tiempo o no necesita leer toda la documentación referente al tema de la búsqueda para solucionar su problema.

Ejemplos de las preguntas que tratan los sistemas de BR son las siguientes: "¿Quién asesinó a Lincoln?" o "¿En qué ciudad

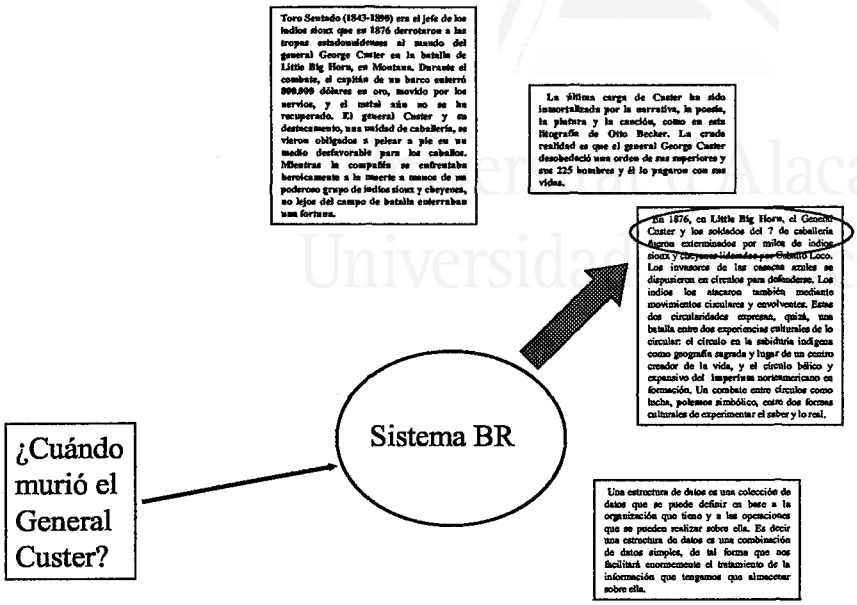


Figura 1.3. Sistema de búsqueda de respuestas

nació Felipe II?". En ambos casos el objetivo del sistema de BR es localizar pequeños fragmentos de texto que contengan la respuesta a la pregunta realizada o incluso, extraer directamente la respuesta buscada.

El interés en la BR por parte de la comunidad científica se ha demostrado por la aparición en 1999, dentro de las conferencias TREC (TREC-8, 1999), de una tarea específica dedicada a la evaluación y comparación de sistemas de BR.

1.1.4 Relación entre los sistemas de RI, BR y EI

Los sistemas de BR y EI van un paso más allá de la mera indicación que hacen los sistemas de RI de la relevancia o no que tiene un documento con respecto a una determinada pregunta. Por ejemplo, si un sistema de BR quiere contestar satisfactoriamente una pregunta del usuario, necesita entender tanto la pregunta como la colección de textos donde puede hallarse la respuesta. Esto obliga

a llevar a cabo una serie de análisis adicionales a los meramente estadísticos que suelen realizar los sistemas de RI. Este análisis supone un incremento notable del coste computacional. Si la colección de textos donde se debe buscar la respuesta es de tamaño considerable y el usuario necesita conocer la solución en tiempo limitado, los sistemas de BR suelen utilizar en primer lugar un sistema de RI, que procesa la pregunta y devuelve una cantidad de texto limitada que previsiblemente contendrá la respuesta (ver figura 1.4). Posteriormente, el sistema de BR realizará su estudio sobre esta parte de la colección.

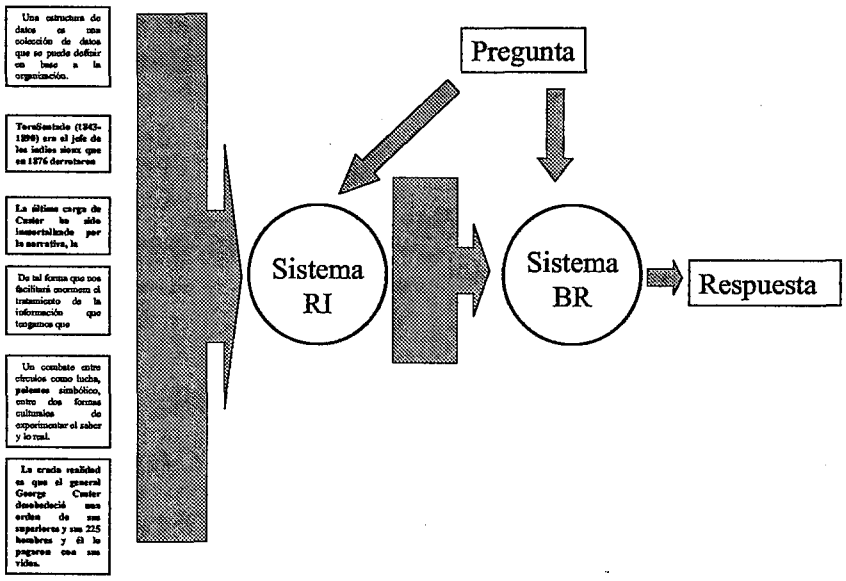


Figura 1.4. Uso previo de un sistema de RI en un sistema de BR

Este mismo análisis se puede aplicar a los sistemas de EI, en su búsqueda de datos para cumplimentar las plantillas previamente fijadas. Es por ello por lo que muchos sistemas de BR o EI aplican un sistema de RI de forma previa que permita reducir la cantidad

de textos sobre los que se aplicarían procesos computacionalmente complejos (Cowie y Lehnert, 1996).

1.2 Los sistemas de Recuperación de Información basados en pasajes

Dentro de los sistemas de RI han aparecido principalmente dos líneas de investigación que tienen como objetivo el de mejorar la eficacia del proceso de recuperación: la aplicación de técnicas de Procesamiento de Lenguaje Natural (en adelante PLN) y la recuperación basada en pasajes (en adelante RP).

La primera incorpora técnicas como las de etiquetado léxico, análisis sintáctico o semántico al proceso de RI. La aplicación de dichas técnicas presenta fundamentalmente el inconveniente del incremento de carga computacional que suponen. Este problema es de gran importancia dentro de lo que es la RI, ya que uno de los objetivos que tienen la mayoría de sistemas de RI, además de recuperar los documentos relevantes, es que esa recuperación se haga en un breve intervalo de tiempo. Además, otro inconveniente es que, hasta ahora, los intentos que se han realizado para incorporar técnicas de PLN, no han mostrado mejoras sustanciales sobre los sistemas estándar (Strzalkowski et al., 1998; Spark-Jones, 1999).

La segunda línea, la RP, se diferencia de lo que sería la RI, basada en el análisis del documento completo (en adelante RID), en la forma de valorar la relevancia de un documento. Estos últimos calculan la relevancia de un documento con respecto a una pregunta mediante la aplicación de una serie de medidas que valoran sobre todo la frecuencia de aparición de los términos de la pregunta en el documento completo. Estas medidas presentan varios inconvenientes, como por ejemplo el hecho de favorecer a los documentos en función de su tamaño.

En contraposición, los modelos de RP estudian la aparición de los términos de la pregunta en fragmentos contiguos de texto dentro de cada documento, a los que se denomina pasajes. Trabajos previos (Callan, 1994; Kaszkiel y Zobel, 1997) demuestran que la utilización de fragmentos de documentos como unidad básica

de información, para calcular la relevancia de un documento con respecto a una pregunta, mejora sensiblemente los resultados de los sistemas de RI. Esta mejora se debe, principalmente, a que ya no sólo se valora el hecho de que los términos de la pregunta aparezcan en el documento, sino que además, se valora la proximidad con que aparecen.

Adicionalmente, los sistemas de RP ofrecen la ventaja de indicar no sólo qué documento es relevante, sino que además, permiten localizar, qué parte del documento es realmente relevante. Este aspecto es de gran utilidad por ejemplo, cuando la salida del sistema de RI es utilizada por un sistema de BR o EI, ya que limita considerablemente la cantidad de texto que estos sistemas deben procesar. Las ventajas de aplicar un sistema de RP en la tarea de BR se muestran en (Vicedo et al., 2002).

1.3 Motivación y objetivos de la tesis

Como ya se ha comentado, los modelos de RP son una propuesta de sistema de RI basada en la división de un documento en una serie de fragmentos de texto, que permiten definir la relevancia de un documento en función de la relevancia de cada uno de esos fragmentos de texto. Sin embargo, no se ha llegado a un consenso acerca de cómo definir esos fragmentos de texto (o pasajes) de forma que el sistema alcance un comportamiento óptimo, ni cómo obtener la relevancia de un documento en función de la relevancia de los pasajes que lo forman.

Para calcular la similitud de un pasaje con respecto a una pregunta, los modelos de RP se basan en medidas de similitud ya utilizadas por sistemas de RI tradicionales como son las medidas del coseno (Salton y Buckley, 1988) y del coseno pivotado (Singhal et al., 1996).

Por otro lado, en función de cómo se aborda la división del documento en pasajes se diferencian tres enfoques de sistemas de RP: modelos basados en el discurso, modelos semánticos y modelos de ventana (Callan, 1994). Los modelos del discurso (Salton et al., 1993; Wilkinson, 1994; Brown y Yule, 1983) utilizan las pro-

piedades de estructura del documento, tales como frases, marcas de párrafo o marcas HTML, para definir los pasajes. Los modelos semánticos (Hearst y Plaunt, 1993; Salton et al., 1996; K. Richmond y Amitay, 1997) se basan en la aparición de tópicos en el documento para definir los pasajes. Los modelos de ventana dividen los documentos en pasajes de tamaño fijo (Callan, 1994; Zobel et al., 1995; Kaszkiel y Zobel, 2001). Para realizar esta división, estos modelos pueden basarse o no en la estructura del documento.

Todos los modelos citados emplean fundamentalmente párrafos y/o palabras como unidad de información básica a partir de la que se definen los pasajes. Los modelos que utilizan el párrafo para definir los pasajes pueden tener problemas en el momento de definir los pasajes si no se dispone de información acerca de la composición de los párrafos en el documento original. Además los párrafos pueden utilizarse en ocasiones, más por motivos visuales que por la propia estructuración del documento. Por otra parte, los modelos basados en el uso de la palabra como unidad para definir los pasajes, son muy dependientes del estilo de escritura utilizado en los documentos. Además, si se utilizan únicamente las palabras como elemento a considerar en la definición del pasaje, puede ocurrir que los pasajes considerados relevantes carezcan de estructura, al poder empezar y finalizar en cualquier parte del documento. Esto dificulta en gran medida la comprensión del texto recuperado.

El trabajo que se presenta a continuación es una nueva propuesta de sistema de RP, que se encuadraría dentro de los modelos de ventana, pero diferenciándose de las actuales propuestas, fundamentalmente, en la unidad que se utiliza tanto para definir los pasajes como en la medida empleada para calcular la similitud de los mismos. Las propuestas realizadas hasta ahora tanto para definir los pasajes como para calcular la similitud se basan en el tamaño del pasaje definido en número de palabras o caracteres que lo forman.

Nuestra línea de investigación ha sido la de utilizar la frase como unidad para la definición y cálculo de similitud de los pasajes. El uso de la frase como unidad en un sistema de RP permite

disponer siempre de pasajes con estructura y sentido, así como independizar el estilo utilizado por los diversos autores de los textos donde se realiza la búsqueda de documentos relevantes.

El objetivo principal perseguido consiste en el diseño y desarrollo de un modelo de RP que utilice las frases como unidad de definición de pasajes. Esto permite construir un modelo de RP muy flexible y que además permite generar pasajes con estructura que puedan ser entendidos por un usuario o tratados por un sistema de BR.

Este modelo también contempla la definición de medidas de similitud que permiten optimizar el rendimiento del sistema en función de la pregunta y de las colecciones de texto utilizadas, e incluso, de la tarea en concreto en la que se emplee.

Un segundo objetivo de este trabajo es el de estudiar y analizar las propuestas de sistemas de RI más importantes, tanto aquellas que se basan en el estudio del documento completo como las de RP.

Un tercer objetivo es el de evaluar el sistema propuesto tanto en las tareas de RI, como en otras tareas en las que pueda ser aplicado como son las de preseleccionar los pasajes más relevantes para sistemas de BR o sistemas de Selección Interactiva de Documentos (en adelante SID). Se pretende que la evaluación sea lo más completa e independiente posible y que además, permita contrastar los resultados obtenidos frente a otras propuestas de sistemas de RI. Por ello, el sistema ha participado de forma directa en las conferencias CLEF y como apoyo de un sistema de BR en las conferencias TREC.

1.4 Organización de la tesis

Esta tesis se ha estructurado en los siguientes capítulos:

Capítulo 2. Los sistemas de Recuperación de Información. Este capítulo presenta un estudio de los sistemas de RI. En primer lugar, se indican las características básicas de los sistemas de RI y los recursos que utilizan. A continuación se introducen

las principales características de los modelos existentes de RI, haciendo especial hincapié en los modelos RID.

Capítulo 3. Los sistemas de Recuperación de Información basados en pasajes. Este capítulo analiza los principales inconvenientes que tienen los sistemas de RID. Se introducen las características de los modelos de RP y las ventajas que aportan sobre los primeros. Se realiza una clasificación de los mismos y se comenta las diferentes aproximaciones existentes.

Capítulo 4. IR-n: Definición del sistema. En este capítulo se presenta el sistema de RP diseñado en base a las propuestas realizadas en esta tesis. A este sistema se le ha denominado IR-n. En primer lugar se comenta los inconvenientes detectados en las propuestas actuales de sistemas de RP y se define un nuevo modelo que pretende paliar estas deficiencias. A continuación se comparan las principales características del modelo IR-n frente al resto de propuestas de modelos de RP. Finalmente, se detallan las principales características a nivel de diseño e implementación del sistema IR-n.

Capítulo 5. Evaluación del sistema IR-n en tareas de Recuperación de Información. En este capítulo se presentan los trabajos de experimentación y evaluación que se han efectuado con el sistema IR-n. En primer lugar se estudian los principales métodos de evaluación de sistemas de RI y se presenta la colección de test que se utilizará para el entrenamiento del sistema IR-n. Esta colección estará formada por las colecciones utilizadas en la edición CLEF-2001 para la evaluación de sistemas de RI. Posteriormente se define el proceso de entrenamiento que tiene como objetivo ajustar las características del modelo IR-n que permitan obtener los resultados óptimos. Para finalizar se presentará la evaluación final del sistema, que se realizó en dos fases. La primera de ellas mediante la participación en la edición CLEF-2002. Esto ha permitido comparar el sistema IR-n con diferentes propuestas de modelos de RI. La segunda de ellas contempla una comparativa

Volver al índice/Tornar a l'índex

del sistema IR-n frente a los sistemas de RP más conocidos.

Capítulo 6. Evaluación del sistema IR-n en otras tareas. En este capítulo se presenta el entrenamiento realizado sobre el sistema IR-n para su adecuación a la realización de las tareas de BR y de Selección Interactiva de Documentos (en adelante SID). En ambos casos se comentan los objetivos de cada una de las tareas, así como la parametrización del sistema para su aplicación a dichas tareas. Dentro de la evaluación se comentará la participación del sistema IR-n junto al sistema de BR denominado SEMQA en la conferencia TREC-10 así como los resultados obtenidos por el sistema IR-n en la tareas de SID en el CLEF-2002.

Capítulo 7. Conclusiones y trabajos futuros. En este capítulo se recogen las conclusiones obtenidas al desarrollar este trabajo y se definen las diferentes líneas de trabajo que se pretenden desarrollar en el futuro.

Finalmente se muestran las referencias bibliográficas utilizadas en el desarrollo de esta tesis.

Apéndice A. Resultados completos de los experimentos realizados. Este apéndice presenta en detalle los resultados de los diferentes experimentos desarrollados en la fase de entrenamiento del sistema, que se han incorporado de forma resumida en el capítulo 5 para facilitar su lectura.

2. Los sistemas de Recuperación de Información

Este capítulo tiene como objetivo el estudiar los aspectos básicos de los sistemas de RI. Este estudio comprende en primer lugar, la descripción de los objetivos a cumplir por estos sistemas. A continuación, se revisarán los principales elementos que utilizan para llevar a cabo sus tareas. Posteriormente se describirán las diferentes propuestas de modelos de RI y las diferentes implementaciones de dichos modelos. Se incidirá en las técnicas principales de expansión de la pregunta y para finalizar se realizará un estudio que define las características más valoradas por los usuarios de los sistemas de RI.

2.1 Conceptos básicos

A un nivel muy abstracto, la RI puede parecer una simple técnica que tiene como objetivo localizar en una colección los documentos que son relevantes a una petición de información realizada por un usuario. No obstante, en la práctica el problema que pretende solucionar es mucho más complejo debido al cambio que se ha producido en el tipo de usuario que suele utilizar estos sistemas. Los primeros sistemas de RI eran manejados por usuarios especializados. Actualmente, debido al desarrollo de la Informática y de las comunicaciones, y especialmente de Internet, es el usuario final el que utiliza estos sistemas, tales como los buscadores de Internet.

Por ello, el objetivo principal de los sistemas de RI es permitir a cualquier usuario formular consultas y devolverle únicamente los documentos que le son relevantes, teniendo en cuenta los diferentes perfiles de usuario existentes. Así, en Baeza-Yates y Ribeiro-

Neto (1999) la RI se define como "Dada una necesidad de información (consulta + perfil de usuario) y una colección de documentos, ordenar los documentos en función de su mayor o menor relevancia para esa necesidad y presentar un subconjunto de los más relevantes". Los conceptos claves que se extraen de esta definición son varios. En primer lugar la idea de necesidad de información, que suele venir representado por una pregunta o consulta como por ejemplo "¿dónde murió el General Custer?" o "quiero información sobre el nacimiento de Felipe II?". En segundo lugar, en la definición se añade el concepto de perfil de usuario, que hace referencia a su mayor o menor conocimiento del funcionamiento interno del sistema de RI. Y en tercer lugar, se extrae la necesidad de puntuar o valorar la relevancia de cada documento con respecto a la consulta formulada, para poder ordenarlos en función de dicha puntuación.

Si se estudia un sistema de RI en función de la información de entrada que recibe y la información de salida que produce, se obtendría que, como entrada, recibe la consulta del usuario y la colección de documentos que dispone para efectuar la búsqueda y, como salida, produce una lista ordenada de documentos relevantes.

Por ejemplo, la entrada del sistema podría ser la indicada en (1)

(1) **Consulta** : La muerte del General Custer

Colección de documentos disponible:

Documento 1: El General Custer murió un domingo de Junio en Little Big Horn.

Documento 2: El general Custer, también conocido como el general de los cabellos rubios fue el general más joven de la historia de los Estados Unidos.

Documento 3 : El general cabellos rubios encontró la muerte, junto con sus hombres en las cercanías del río Little Big Horn.

Documento 4 : El oeste americano ha inspirado a numerosos autores.

La salida que produciría un sistema de RI ante dicha consulta sería la lista de documentos relevantes, en este caso los documentos 1 y 3.

Los buscadores de Internet, además de mostrar la dirección web del documento considerado como relevante, muestran un pequeño contenido del mismo para facilitar al usuario comprobar si son de su interés sin tener que acceder a los mismos. Un ejemplo de esta información se puede ver en la figura 2.1.

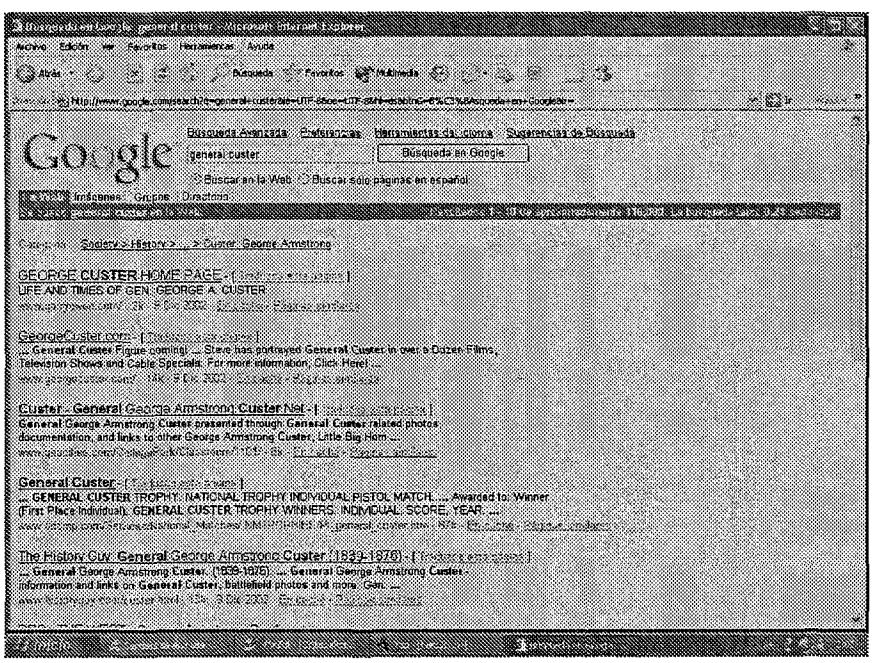


Figura 2.1. Ejemplo de buscador de Internet

Para determinar la relevancia del documento con respecto a una pregunta, los sistemas de RI suelen valorar fundamentalmente la cantidad de términos que ambos comparten. Además, esa relevancia se cuantifica según una serie de aspectos que indican la importancia de esos términos que comparten la pregunta y el documento y que definen lo que se denomina el modelo de similitud del sistema de RI. Dichos aspectos son:

- **La frecuencia de aparición de los términos de la pregunta en el documento.** Este aspecto mide la cantidad de apariciones de cada uno de los términos de la pregunta dentro del documento. Así, a mayor número de apariciones de los términos de la pregunta en el documento, éste tendrá una mayor valoración.
- **La frecuencia de aparición del término en la colección de documentos.** Este aspecto, también denominado "peso" o "valor del término" indica la capacidad de "discriminación" de los términos en el momento de determinar la relevancia de un documento con respecto a la pregunta. Éste es un aspecto muy importante y que diferencia muchas propuestas de modelos de RI por la forma en la que se valora cada término. La propuesta generalmente aceptada es la que indica que cuanto más raro en la colección es un término más peso o valor discriminatorio tiene (Salton y McGill, 1983). De hecho, algunos sistemas proponen que si la frecuencia de aparición de un término en una colección supera un determinado umbral, no debe tenerse en cuenta cuando se realizan los cálculos de similitud, ya que se considera que no permite discriminar entre un documento relevante de uno que no lo es.
  La técnica de asignación de pesos más utilizada es la desarrollada en (Sparck-Jones, 1972) donde a cada término se le asigna un peso calculado en función del valor inverso de su frecuencia de aparición en el conjunto de documentos de la colección (*inverse document frequency - idf*). Otros aspectos que también se pueden tener en cuenta cuando se mide este valor puede ser su categoría gramatical. Los nombres y verbos suelen tener mayor valoración que adjetivos y adverbios, e incluso hay algunos

tipos de términos que ni siquiera se tienen en consideración, como pueden ser, entre otros, las preposiciones (Henstock et al., 2001).

- **El tamaño del documento.** Es lógico pensar que un documento de mayor tamaño tiene una probabilidad mayor de contener los términos de la pregunta que un documento de menor tamaño. Para poder comparar la relevancia de documentos de diferente tamaño se suelen aplicar medidas de normalización. Estas medidas ajustan el valor de relevancia del documento en función del tamaño del mismo. El tamaño de los documentos se suelen medir en número de palabras o bytes que lo forman.

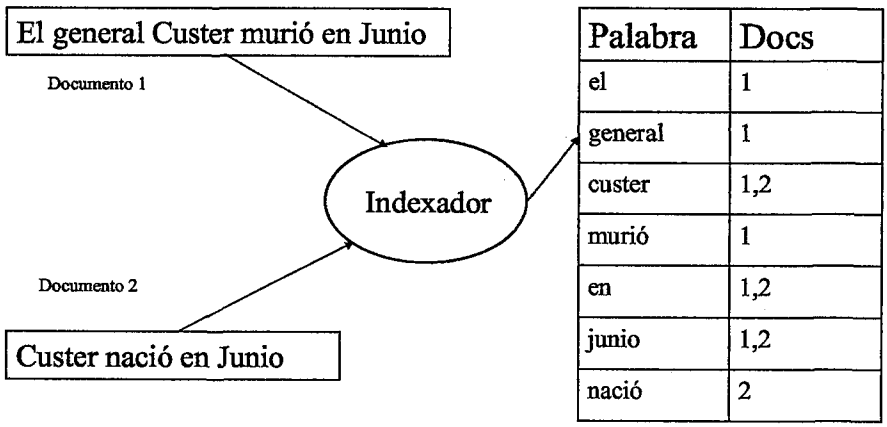
Considerando estos aspectos, el proceso más sencillo para determinar cuál de los documentos del ejemplo es más relevante, consistiría en recorrer secuencialmente cada uno de ellos, contando el número de veces que aparecen las palabras de la pregunta en ellos. Si la colección de documentos donde se debe efectuar la búsqueda es pequeña, el proceso no tarda mucho tiempo en realizarse. Pero a medida que la colección de documentos crece, el tiempo necesario para realizar este proceso se incrementa de tal forma, que aleja al sistema de RI del requerimiento del tiempo de respuesta.

Para solucionar este problema aparece el concepto de índice. La idea que se tiene de un índice es el de un elemento que nos facilita una búsqueda. Por ejemplo, el índice de un libro permite conocer en qué páginas se encuentra detallado determinado tema, sin tener que leer todo el libro. De forma muy similar, un índice en una Base de Datos permite acelerar el proceso de búsqueda de información.

Dentro del ámbito de la RI, el punto de partida del proceso es la pregunta o consulta del usuario. Ésta a su vez está formada por una serie de palabras. La idea de índice dentro de la RI, se utiliza para conocer en qué documentos se encuentran dichas palabras e incluso, en qué posiciones del documento.

Dado que la palabra es el elemento que define la forma de acceder, ésta es la clave de acceso. Así, para cada palabra se ha de almacenar básicamente en qué documentos aparece, formando una

estructura, denominada índice invertido, como la que se presenta en la figura 2.2. En esta figura se muestra el proceso de indexación o construcción de un índice, este proceso parte de los documentos y genera información para cada una de las palabras contenidas en los mismos.



Palabra	Docs
el	1
general	1
custer	1,2
murió	1
en	1,2
junio	1,2
nació	2

Figura 2.2. Ejemplo de indexación

Una vez construido el índice, el proceso de búsqueda de documentos relevantes, consiste básicamente en acceder al índice para conocer en qué documentos se hallan las palabras que forman la pregunta. Una vez realizado este proceso, se efectúan los cálculos en función del modelo de recuperación que utilice el sistema de RI y se muestra una lista ordenada en función de la relevancia calculada para los documentos (ver figura 2.3).

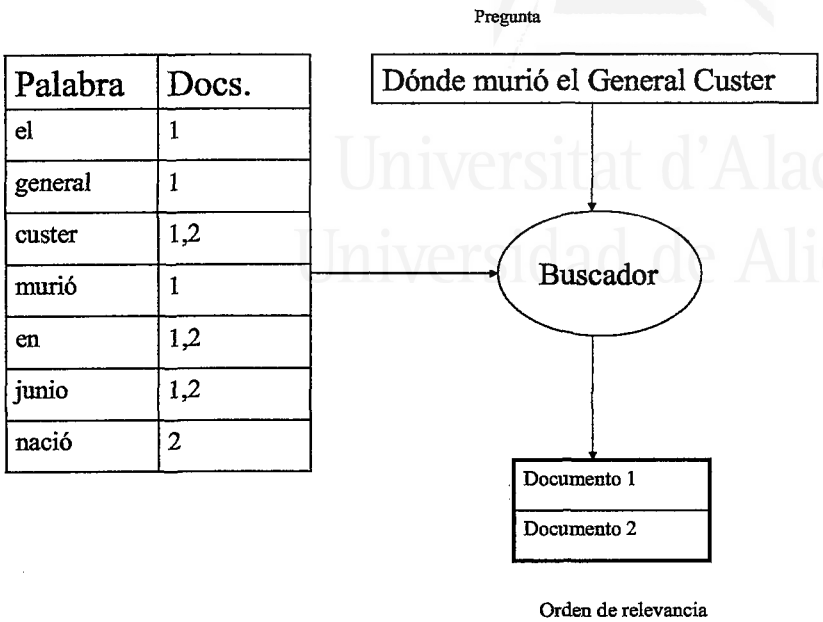


Figura 2.3. Ejemplo de búsqueda

2.2 Arquitectura general de un sistema de RI

Implícitamente se han definido las dos tareas principales de un sistema de RI: las de indexación y búsqueda. No obstante, aparece otro concepto que es la definición del modelo de búsqueda, es decir, qué técnicas va a utilizar el sistema de RI en función del conocimiento del que dispone sobre la aparición de las palabras de la pregunta en los documentos para determinar la relevancia de éstos o bien el orden de relevancia de los documentos disponibles.

Esto nos lleva a una visión del sistema de RI que puede resumirse en la figura 2.4. En esta figura pueden verse los dos usuarios principales del sistema de RI. En primer lugar el usuario administrador del sistema que debe indicar al mismo qué modelo de recuperación se debe utilizar, así como la colección de documentos donde se deberá buscar la información. En segundo lugar, el usuario final del sistema que establece una comunicación suminis-

trando una pregunta o tema de interés, para el cual el sistema de RI deberá devolver una serie de documentos ordenados en función de la relevancia a la pregunta.

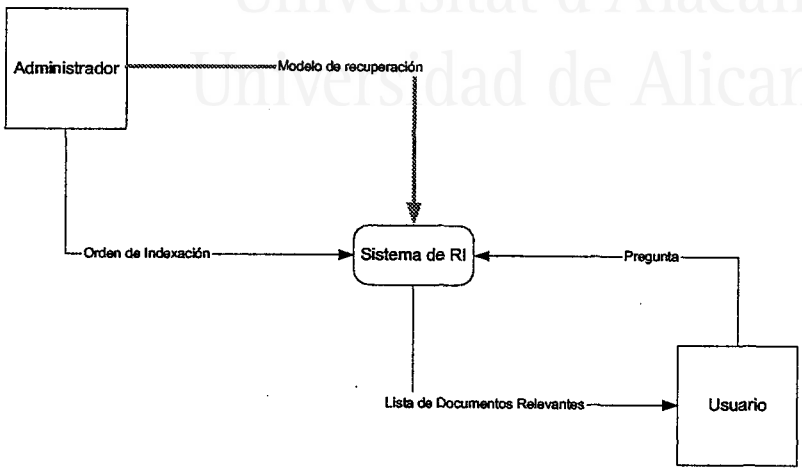


Figura 2.4. Diagrama conceptual de un sistema de RI

No obstante, el sistema de RI necesita de otros procesos complementarios para realizar su tarea. En primer lugar, tanto los documentos como las preguntas pueden requerir un proceso previo, ya que es posible que toda la información que contenga el documento o la pregunta no deba ser considerada en los procesos de búsqueda de información. También es destacable el hecho de que parte de los elementos que se deben considerar en los procesos de indexación y búsqueda no tienen porqué ser las palabras completas, sino una parte de las mismas. Además, hay que considerar el hecho de que los sistemas de RI incorporan técnicas de procesamiento o expansión de la pregunta original, modificándola de forma que permitan mejorar los resultados.

Los principales módulos en los que se descompone un sistema de RI son los que se indican a continuación, los cuales se relacionan tal y como se muestra en la figura 2.5.

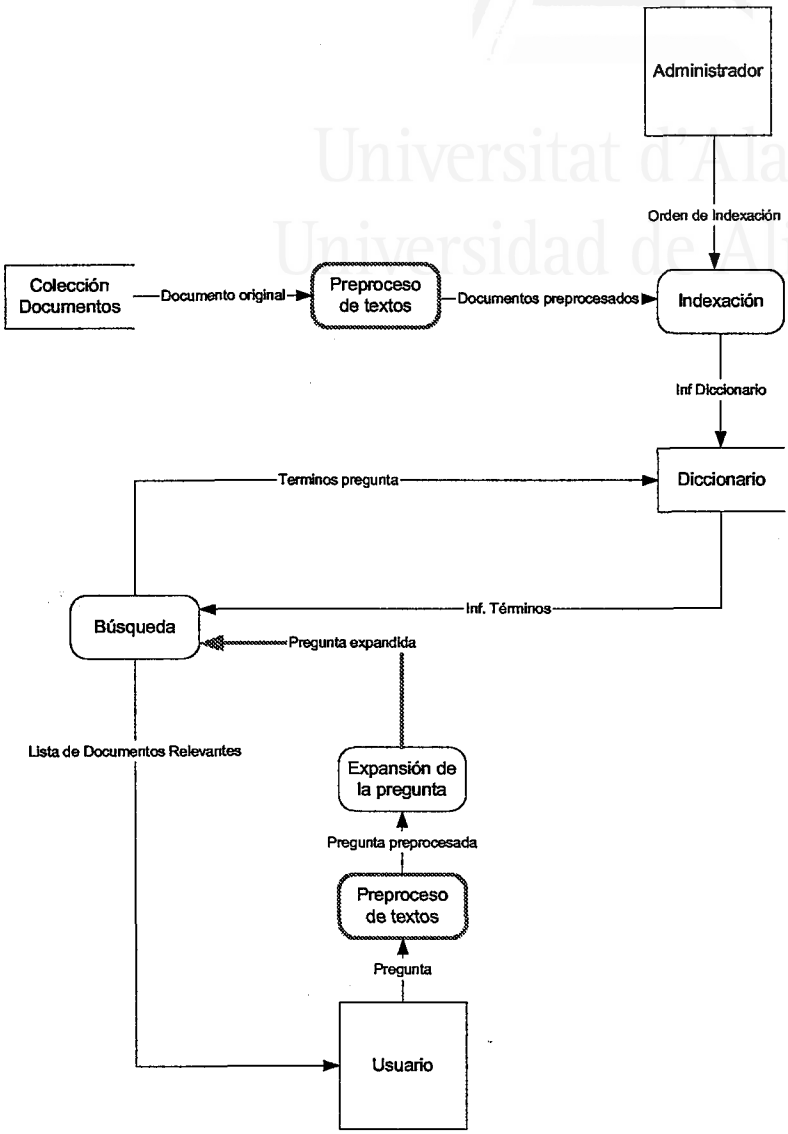


Figura 2.5. Principales módulos de un sistema de RI

- **Módulo de preproceso de textos.** Tiene como objetivo discriminar la información que no se va a utilizar, así como determinar cuál es la parte de cada palabra que debe ser utilizada para la indexación y búsqueda. Este proceso se aplica tanto a documentos como a preguntas.
- **Módulo de indexación.** Tiene como objetivo fundamental el de organizar la información contenida en la colección de documentos de forma que facilite el acceso a la misma en el proceso de búsqueda.
- **Módulo de gestión de la pregunta.** Este módulo tiene como objetivo preparar la pregunta para que pueda ser utilizada por el módulo de búsqueda. Puede suponer un simple tratamiento de preproceso de la pregunta tal como se efectuó sobre los documentos, o una incorporación de nuevos términos a la pregunta original con el objeto de mejorar los resultados (como se verá en el apartado referido a expansión de la pregunta).
- **Módulo de búsqueda.** Tiene como objetivo recibir la pregunta y devolver los documentos relevantes a la misma. Para ello, deberá basarse en las estructuras generadas por el módulo de indexación, así como en el modelo de similitud utilizado.

Los sistemas de RI se diferencian entre sí por las diferentes estrategias que emplean en cada uno de los módulos. En los siguientes apartados se describirán las principales características de cada uno de ellos.

2.2.1 Módulo de preproceso de textos

Todos los sistemas de RI utilizan una serie de elementos o utilidades que facilitan enormemente el proceso de indexación y búsqueda de información. En (Baeza-Yates y Ribeiro-Neto, 1999) se cita las fundamentales: *tokenización*, *palabras de parada* y *lematización* (o *stemming*).

A estas etapas cabría añadir una más referente al tratamiento de posibles errores en los documentos originales, como puede ser la existencia de palabras excesivamente largas en los mismos. Este proceso se denomina *filtrado de palabras adicionales.*

Tokenización. Uno de los objetivos del la tokenización es convertir un conjunto de caracteres (un documento) en un conjunto de palabras o términos. Estas palabras o términos son las candidatas a ser los elementos a indexar.

Este proceso localiza en el texto todos aquellos caracteres que pueden delimitar una palabra, tales como un espacio en blanco, un punto, una coma, etc. Hay que tener especial cuidado, ya que hay signos de puntuación que forman parte de la palabra en sí, como puede ser "s.a.". También hay que considerar dentro de este apartado los símbolos de puntuación que al eliminarlos permiten la descomposición de palabras, como es el caso de "palestino-israelí", al quitar el guión se generan dos palabras "palestino" e "israelí".

Otro objetivo es el de uniformizar las palabras convirtiéndolas en mayúsculas o minúsculas según se considere.

También se debe considerar el tratamiento de los números. En primer lugar reconocer qué palabras son números (1989, 2034...) e incluso puede plantearse, convertir números escritos en letras a números representados mediante dígitos (por ejemplo de "cincuenta mil" a "50000"). En segundo lugar determinar si dichos números son elementos a considerar. Hay sistemas que no consideran los números como elementos a indexar y otros que sí, dependiendo fundamentalmente del tipo de documentos y/o aplicación que se esté utilizando. Un ejemplo de este tratamiento es el definido en (2)

(2) **Entrada:**
El proceso de paz palestino-israelí se llevó a cabo en secreto.
EFE19943746834759342kkj495

Salida:
el proceso de paz palestino israelí se llevó a cabo en

secreto
efe19943746834759342kkj495

En dicho ejemplo se ha eliminado el símbolo '-' de la palabra "palestino-israelí", generando las palabras "palestino" e "israelí". Además, ha quitado el punto final de la frase, que se hallaba en la palabra "secreto". Finalmente se han convertido todas las mayúsculas en minúsculas.

Palabras de parada. Una *palabra de parada* (*stopword*) es aquella palabra que puede aparecer en la pregunta y no se considera importante su aparición en el documento para determinar si éste es relevante o no.

Para determinar qué palabra es una *palabra de parada*, se puede valorar el porcentaje de documentos de la colección en la que una palabra aparece. Si una palabra se encuentra en un porcentaje elevado de documentos se suele considerar como *palabra de parada*, ya que difícilmente se podrá utilizar para discriminar entre varios documentos. Los artículos, preposiciones, conjunciones, e incluso algunos verbos, adjetivos y adverbios suelen formar parte de las listas de *palabras de parada* que utilizan los sistemas de RI. Los siguientes términos en inglés constituyen algunos ejemplos: "he", "it", "to" y "the".

El conjunto de estas palabras se denomina *lista de palabras de parada*. Existen varias de estas listas que se han obtenido en estudios específicos a tal efecto (Fox, 1992; Rijsbergen, 1979; Buckley et al., 1994). Estas listas se generan seleccionando las palabras más frecuentes en las colecciones de documentos utilizadas y, en algunos casos, posteriormente se añaden de forma manual una serie de palabras también consideradas irrelevantes.

Dado que estas palabras no tienen poder de discriminación para determinar si un documento es más relevante que otro, no se indexan, con lo que se consigue reducir notablemente el tamaño de los índices y acelerar el proceso de recuperación. En experimentos realizados sobre diferentes colecciones del TREC, el uso de *palabras de parada* permitía disminuir en un 25% el tamaño de los índices generados (Witten et al., 1999).

A pesar del beneficio que supone no indexar estas palabras, puede ocurrir que su eliminación impida localizar algunos documentos relevantes. Un ejemplo de este hecho sería si un usuario buscara documentos que contuviesen la frase "to be or not to be" que está formado completamente por *palabras de parada.*

En contraposición, aquellas palabras que no aparecen en la lista de *palabras de parada,* se consideran lo suficientemente discriminantes como para representar el contenido de un documento y por tanto, son indexables. Estos términos reciben la denominación de *palabras clave* (*keywords*).

En (3) se muestra como quedaría el texto descrito en (2) al eliminar las *palabras de parada.*

(3) **Entrada**:
el proceso de paz palestino israelí se llevó a cabo en secreto
efe19943746834759342kkj495
Salida:
proceso paz palestino israelí llevó cabo secreto
efe19943746834759342kkj495

En este ejemplo se puede comprobar que se han quitado del documento de salida las palabras "el", "de", "se", "a" y "en".

Lematización y/o stemming. Uno de los mayores problemas con los que se encuentra la RI es la utilización de diferentes palabras para expresar la misma idea en documentos y preguntas. En (4) se muestra uno de estos casos.

(4) **Pregunta** : ¿Quién asesinó a Marat?
Documento : La asesina de Marat fue la realista Carlota Corday el 13 de julio. Ella lo acuchilló en la bañera.

Como se puede observar, los términos "asesinó" y "asesina" se refieren al mismo hecho. No obstante, si el proceso de indexación se realiza empleando directamente el término, posiblemente el documento no se consideraría relevante, ya que documento y pregunta únicamente comparten uno de los términos.

Para solucionar este problema la mayoría de sistemas de RI, utilizan términos que no corresponden exactamente con la palabra. Estos términos pueden ser directamente el *lema* de la palabra o bien el *stem*.

El *lema* es la raíz de la palabra, mientras el *stem* es una porción de una palabra, a la cual se han eliminado los afijos, bien sufijos y/o prefijos. El tipo de idioma es el que suele definir que término hay que utilizar para indexar. Mientras en inglés se suele utilizar el stem obtenido a través de algoritmos como el de Porter (Porter, 1980), para español se utilizan variaciones del mismo (C. Figuerola y Berrocal, 2001) o el lema (J. Broglio y Nachbar, 1994).

En (5) se puede ver un ejemplo de la salida que se obtendría al utilizar un *stemmer*.

(5) **Entrada** :
proceso paz palestino israelí llevó cabo secreto efe19943746834759342kkj495
Salida :
proces paz palestin israeli llev cabo secret efe19943746834759342kkj495

Filtrado de palabras adicionales. Existen palabras que pueden ser eliminadas debido a que no son consideradas interesantes

para la determinación de la relevancia de los documentos de la colección.

Ejemplos de estas palabras son aquellas que tienen una excesiva longitud, que puede deberse a algún error que contenga el documento original. Las palabras de una longitud mayor de 20 son en general errores del documento (Moffat y Zobel, 1996), debido a que por algún motivo aparecen unidas, o son marcas que pueden definir el documento.

Además, en función de la colección utilizada, los números (o los superiores a alguna cifra) pueden no considerarse interesantes para ser indexados (Baeza-Yates y Ribeiro-Neto, 1999).

En (6) se puede ver la salida que se obtendría a partir del documento generado en la fase anterior al aplicar el filtrado.

(6) **Entrada** :
proces paz palestin israeli llev cabo secret
efe19943746834759342kkj495
Salida :
proces paz palestin israeli llev cabo secret

En este caso se ha eliminado de la frase original la palabra "efe19943746834759342kkj495", debido que su tamaño es superior a 20 caracteres. Así, la salida del módulo de preproceso de textos, y por tanto, el conjunto de términos a indexar del ejemplo anterior serían las siguientes: "proces", "paz", "palestin", "israeli", "llev", "cabo" y "secret".

2.2.2 Módulo de indexación

Intentar localizar información en un libro que no dispone de índice y/o glosario de términos, puede suponer un elevado coste de tiempo, ya que implica tener que realizar una búsqueda secuencial sobre el contenido del libro. La misma circunstancia ocurre cuando se intenta localizar documentos relevantes a una determinada

consulta si no se dispone de ninguna estructura que facilite el acceso.

La indexación consiste en la generación de una serie de estructuras que faciliten la localización de los documentos que contienen un término en concreto. En Witten et al. (1999) se define como objetivo principal del proceso de indexación el de generar un *vocabulario* que contenga información al respecto de los diferentes términos y los documentos en los que aparecen. También se debe generar una estructura que contenga información sobre el tamaño de cada uno de los documentos de la colección (*fichero de documentos*), en el caso de que el modelo de similitud a utilizar tenga en cuenta este valor.

Así, la indexación consiste en la generación de estructuras que faciliten la localización de los documentos en función de las palabras que contengan. Estas estructuras son dos: el vocabulario y el fichero de documentos.

Así dado el ejemplo de documentos descrito en (7)

(7) Documento 1
El proceso de paz palestino-israelí se llevo a cabo en secreto. Sharon habló con Arafat.

Documento 2
El líder israelí Sharon habló de paz con el líder palestino Arafat.

Tras el preproceso de ambos documentos se generarían las dos listas de términos a indexar, una para cada uno de los documentos, según se muestra en (2.1)

Una información que debe contener el vocabulario es el término, el número de documentos en los que aparece y el nombre de los mismos. En la tabla 2.2 se puede ver una forma de representar esta información.

Documento 1	Documento 2
proces	lider
paz	israeli
palestin	sharon
israeli	habl
llev	paz
cabo	lider
secret	palestin
sharon	arafat
habl	
arafat	

Tabla 2.1. Ejemplo de lista de términos por documento

Término	Num Apariciones	Num Doc	Num Apariciones
arafat	2	1	1
		2	1
cabo	1	1	1
habl	2	1	1
		2	1
israeli	2	1	1
		2	1
lider	1	2	2
llev	1	1	1
palestin	2	1	1
		2	1
paz	2	1	1
		2	1
proces	1	1	1
sharon	2	1	1
		2	1
secret	1	1	1

Tabla 2.2. Ejemplo de vocabulario

La estructura que contiene información de los documentos podría ser la reflejada en la tabla 2.3. En ella se puede comprobar que se ha generado una entrada para cada uno de los dos documentos de la colección. Para cada uno de estos documentos se ha almacenado el nombre del mismo y el número de palabras que contiene.

Nombre	Tamaño
Doc. 1	10
Doc. 2	8

Tabla 2.3. Ejemplo de fichero de documentos

Las estructuras descritas en este apartado son una simplificación de las comúnmente utilizadas. Hay que considerar que en función del modelo de recuperación utilizado se puede requerir más cantidad de información o incluso puede que alguna de la especificada no sea necesaria. También cabe citar que muchos sistemas utilizan técnicas de compresión de la información que se indexa con el objeto de reducir los costes de almacenamiento y transmisión de la información.

2.2.3 Módulo de gestión de la pregunta

Este módulo aplica a la pregunta el mismo preproceso que se aplicó a los documentos de la colección. Esto es coherente, ya que no tiene sentido considerar en las preguntas, las *palabras de parada* eliminadas en el preproceso de la colección de documentos. Tampoco tiene sentido utilizar tipos de términos diferentes, por ejemplo, generar el *lema* como unidad en las preguntas cuando ha sido utilizado el *stem* en los documentos.

Por otra parte, el módulo puede emplear las denominadas *técnicas de expansión de la pregunta*, que incorporan nuevos términos a la pregunta original, con el objeto de mejorar los resultados y posibilitar la localización de documentos que aún siendo relevantes no contienen los términos exactos de la pregunta. Estas técnicas se comentarán en la sección 2.4 de este capítulo.

2.2.4 Módulo de búsqueda

El módulo de búsqueda tiene como objetivo ordenar los documentos en función de la relevancia que tienen hacia una pregunta o tema determinado. Este módulo debe realizar fundamentalmente dos tareas:

1. **Selección de documentos.** Esta tarea debe recuperar, para cada término de la pregunta, aquella información referente a los documentos en los que aparece. Esta información se encuentra disponible en las estructuras creadas en el proceso de indexación.

2. **Cálculo de la relevancia de los documentos.** Esta tarea finalmente debe aplicar una medida de similitud en función de la información que se dispone sobre los términos de la pregunta y sus apariciones en los documentos de la colección.

Mientras que la forma de realizar la primera tarea suele ser muy similar en todos los modelos de RI, la segunda tarea es la que los diferencia. Existen diferentes modelos que determinan la relevancia de los documentos. En el apartado siguiente se comentarán los principales modelos de RI, profundizando en el modo en que realizan el cálculo de relevancia.

2.3 Modelos de Recuperación de Información

Dada una pregunta y una base de datos documental es importante disponer de algún modelo que permita indicar si un documento es relevante o no con respecto a dicha pregunta. Se han propuesto muchos modelos de RI, aunque los tres modelos más conocidos son el modelo lógico o booleano, el modelo del espacio vectorial y el modelo probabilístico. Estos modelos son los más referenciados y utilizados como referencia en diversas comparativas de sistemas de RI.

2.3.1 Modelo lógico o booleano

El modelo lógico es el más simple de los tres citados, y fue el modelo inicialmente utilizado en los sistemas bibliográficos. Se basa en la aplicación de las teorías de conjuntos y del álgebra de Boole, operadores lógicos AND (Y lógico), OR (O lógico) y NOT (negación), a los términos de la pregunta. Una vez definida la pregunta del usuario formando una expresión lógica, el proceso de búsqueda debe determinar si los documentos de la colección la satisfacen, en cuyo caso el documento es considerado relevante.

La mayor ventaja de este sistema es la simplicidad de su desarrollo y su rapidez tanto en los procesos de indexación como de búsqueda. Puede ser adecuado para usuarios expertos que tengan facilidad para convertir sus preguntas en expresiones lógicas, o

tengan experiencia en consultas a Bases de Datos. Además, es un sistema que puede ser utilizado en combinación con otros modelos para excluir documentos no relevantes.

Por otra parte, sus mayores inconvenientes residen en la complejidad que puede suponer trasladar una necesidad de información a una expresión lógica y, sobre todo, en su estrategia de selección de documentos ya que, según la aplicación del modelo lógico, un documento es o no relevante con respecto a una pregunta y no existe ninguna ordenación final de los documentos.

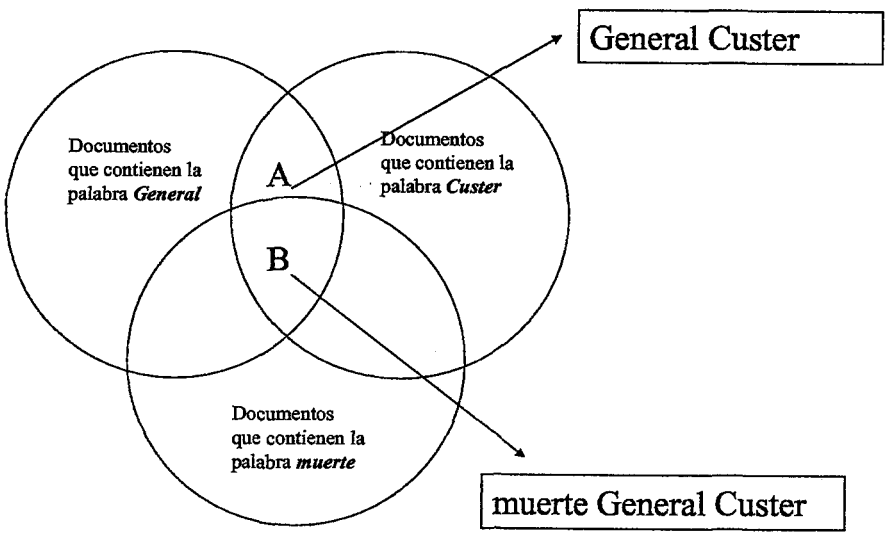


Figura 2.6. Modelo lógico o booleano

En la figura 2.6 se puede ver un ejemplo de funcionamiento del sistema lógico. Ante una búsqueda del tipo "General Custer", el sistema devolvería como relevantes todos aquellos documentos que contuviesen al menos una vez las palabras "General" y "Custer". Si se añadiese a la pregunta la palabra "muerte", sólo se considerarían relevantes los documentos que contuviesen las tres palabras.

En Salton et al. (1983) se presentó una modificación al modelo booleano básico, en la que se añaden los conceptos de valor

o peso de los términos, así como la idea de localización parcial de los términos de la pregunta. Esta propuesta se ha denominado modelo *booleano extendido*, el cual permite ordenar los documentos en función de su relevancia a una pregunta. En este modelo ya no se consideran relevantes todos los documentos que cumplen la expresión lógica en la que se convierte la pregunta, sino que se establece una valoración de los documentos en función de la parte de la expresión lógica que cumplen. Este modelo no ha sido demasiado utilizado, aunque algunos investigadores (Baeza-Yates y Ribeiro-Neto, 1999) consideran que dadas sus características puede tener importancia en el futuro.

2.3.2 Modelo vectorial

El modelo lógico presenta algunos inconvenientes. En primer lugar, cabe considerar que no es necesario que todos los términos de la pregunta se hallen en el documento para que éste pueda ser relevante (Witten et al., 1999). En segundo lugar, tampoco parece tener sentido que todos los términos de la pregunta tengan el mismo peso a la hora de determinar si un documento es relevante o no.

El modelo del espacio vectorial (Salton y McGill, 1983) parte de la base de que la valoración binaria realizada por el modelo lógico no es la más adecuada. La primera aportación del modelo es la asignación de una serie de pesos o valores a cada uno de los términos que aparecen en la colección de documentos.

Esto permite solucionar algunos de los problemas citados del modelo lógico. Primero, cambia la idea de relevancia o no relevancia propia del modelo lógico por la idea de rango de relevancia, es decir, un documento puede ser más o menos relevante. Esto permite dirigir al usuario a que se centre inicialmente en el estudio de los documentos más relevantes en vez de buscar aleatoriamente en un conjunto de documentos relevantes.

Uno de los aspectos que se consideran para definir ese grado de relevancia es el peso o valor del término de la pregunta que aparece en el documento. El establecimiento del peso de cada término es un aspecto considerado de gran importancia para los sistemas de

RI modernos (Buckley, 1994). Para valorar el peso de un término dentro de un documento se deben cuantificar dos aspectos, el primero de ellos es la cantidad de veces que aparece el término en el documento (*frecuencia del término en el documento*). A mayor cantidad de apariciones del término mayor peso tendrá el término en el documento. El segundo aspecto es que se debe cuantificar el número de documentos de la colección en los que aparece dicho término en la colección (*frecuencia del término en la colección*). A mayor cantidad de documentos en los que aparece, menor importancia tendrá dicho término para diferenciar la similitud de un documento con respecto a otro.

De esta forma, cada documento será valorado en función de los términos de la pregunta que contenga y de los valores de dichos términos, no siendo necesario que incluya todos los términos de la pregunta para considerarse relevante.

Así, el modelo vectorial constituye el modelo de representación más utilizado en sistemas de RI debido a su simplicidad y a su buen comportamiento respecto de otras aproximaciones (Salton y McGill, 1983). Su eficiencia ha quedado demostrada en estudios como los de Salton (1989). De hecho, se ha convertido en el sistema de referencia con el que se comparan todos los demás modelos.

La base teórica del modelo vectorial representa en forma de vectores ponderados tanto los documentos como las preguntas, dentro de un espacio n-dimensional. En este espacio, n representa el número de términos indexables.

Así las preguntas $\vec{p}_j$ y los documentos $\vec{d}_i$, se representan mediante vectores de la forma:

$$\vec{p}_j = (pp_{j1}, pp_{j2}, ..., pp_{jn}) \tag{2.1}$$

$$\vec{d}_i = (pd_{i1}, pd_{i2}, ..., pd_{in}) \tag{2.2}$$

Donde pp_{it} y pd_{it} representa el peso asociado al término t en preguntas y documentos.

Al representar tanto la pregunta como el documento en forma de vectores, el modelo calcula la similitud en función de la correlación o cercanía de ambos vectores (ver figura 2.7). En ella, se representa el documento como el vector D y la pregunta como el

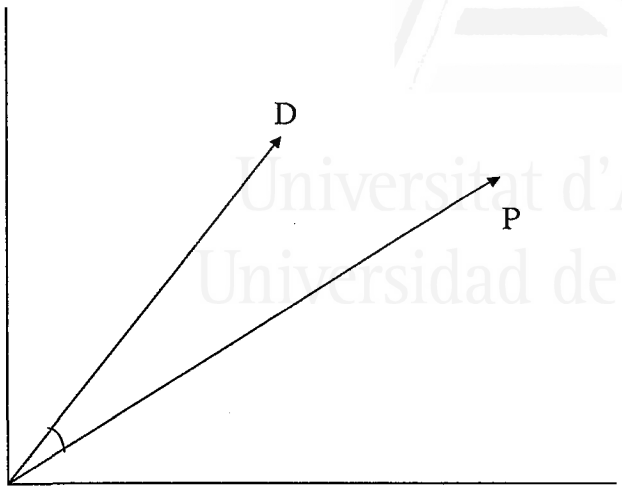


Figura 2.7. Modelo vectorial

vector P. El coseno que forman ambos vectores define la similitud entre los mismos, dado que el coseno toma valores mayores a medida que ambos vectores se encuentran más cerca uno del otro.

Otro aspecto destacable en el proceso de cálculo de similitud propuesto por este modelo es que valora el tamaño de los vectores representados. El tamaño de un documento se cuantifica en lo que se denomina *factor de normalización de un documento*. Este factor es un valor que permitirá comparar entre sí los valores de relevancia de documentos de diversos tamaños. Es evidente que cuanto mayor es el tamaño de un documento, más probabilidad tendrá de contener los términos de una pregunta, sin que ello asegure que sea más relevante. Por ello, este factor valora el tamaño de un documento, de forma que a mayor tamaño el valor de similitud total será menor.

Las ventajas del modelo del espacio vectorial son su simplicidad, sus mejores resultados comparados con los obtenidos por el modelo lógico (Salton, 1989) y, sobre todo, que permite ordenar los documentos en función de un valor que mide su similitud a la pregunta recuperando documentos que se aproximan a la pregunta aunque no contengan todos los términos de la misma.

Como principal inconveniente de este modelo cabe citar que no tiene en cuenta las interdepencias entre términos cuando se les asigna un peso (Baeza-Yates y Ribeiro-Neto, 1999).

El modelo *vectorial* define una forma de representar la información (documentos y preguntas), aunque para calcular la similitud cuantificada entre ellos se ha de acudir a una serie de medidas de similitud. Estas medidas permiten calificar con una puntuación el grado de similitud entre una pregunta y un documento. Una vez obtenida la puntuación para los documentos de una colección, ordenarlos por la relevancia a la pregunta es inmediato. Una medida muy efectiva basada en el modelo del espacio vectorial es la medida del *coseno*. Dentro del mismo modelo, una variante con respecto a la anterior es la del *coseno pivotado*. Estas dos medidas se detallan a continuación.

Medida del coseno. La medida del coseno define una serie de características que contienen las funciones de similitud más efectivas. Incrementa el valor de los documentos que contienen muchos de los términos de la pregunta y cada término aporta un mayor valor en función de la rareza del mismo en la colección. En Salton y Buckley (1988) se definen varias formas de determinar los pesos, pero la medida más conocida para calcular la similitud entre una pregunta q y un documento d es la siguiente (Kaszkiel et al., 1999):

$$sim(q,d) = \frac{\sum_{t \in q \wedge d}(w_{q,t} \cdot w_{d,t})}{W_d \cdot W_q} \tag{2.3}$$

siendo:

$$w_{d,t} = \log_e(f_{d,t} + 1) \tag{2.4}$$

$$w_{q,t} = \log_e(f_{q,t} + 1) \cdot \log_e(\frac{N}{f_t} + 1) \tag{2.5}$$

$$W_d = \sqrt{\sum_{t \in d} w_{d,t}^2} \tag{2.6}$$

$$W_q = \sqrt{\sum_{t \in q} w_{q,t}^2} \tag{2.7}$$

donde:

$f_{x,t}$ es el número de apariciones o frecuencia del término t en la pregunta o documento x.

N es el número de documentos de la colección.

f_t es el número de documentos diferentes que contienen el término t.

La expresión $\log_e(\frac{N}{f_t} + 1)$ es la frecuencia inversa del documento (*inverse document frequency, idf*), o sea el valor del término t en la colección.

$w_{x,t}$ es el peso o poder discriminatorio del término t en la pregunta o en el documento x.

W_x es la representación de la longitud de x, o sea los factores de normalización de la pregunta y el documento.

Medida del coseno pivotado. La medida del coseno, siendo una medida muy sencilla, ofrece una eficacia notable. No obstante, según pruebas realizadas en Singhal et al. (1996) y comentadas en Kaszkiel y Zobel (1997), la medida del coseno favorece la recuperación de documentos cortos debido a los altos valores del factor normalización del documento que se aplica a los documentos con mayor número de palabras. Así a medida que un documento es de mayor tamaño, estos factores de normalización (que se aplican según este tamaño) disminuyen notablemente sus valores de similitud.

La idea básica de la medida del coseno pivotado es definir un nivel a partir del cual, conforme el documento crece en tamaño, el valor de normalización baja con respecto al definido en el coseno, y viceversa. El modelo del coseno pivotado demuestra sus beneficios sobre el modelo del coseno en Singhal et al. (1998). La formulación del modelo del coseno pivotado para el cálculo de similitud entre la pregunta q y el documento d se define como:

$$sim(q,d) = \sum_{t \in q \wedge d} \frac{(w_{q,t} \cdot w_{d,t})}{W_d} \tag{2.8}$$

siendo:

$$w_{q,t} = 1 + \log_e(1 + f_{q,t}) \cdot \log_e(\frac{N+1}{f_t}) \tag{2.9}$$

$$w_{d,t} = 1 + \log_e(f_{d,t} + 1) \tag{2.10}$$

$$W_d = (1 - slope) + slope \cdot \frac{d_{len}}{avr_{-d_{len}}} \tag{2.11}$$

donde:

d_{len} es la longitud del documento medida en bytes.

$avr_{-d_{len}}$ es la media de longitud de los documentos de la colección.

slope es el valor de normalización del coseno. En Singhal et al. (1996) se define que el valor recomendado para obtener una mayor eficacia es el de $0,2$.

2.3.3 Modelo probabilístico

En RI los modelos probabilísticos (Robertson y Jones, 1976; Rijsbergen, 1979) ordenan los documentos de una colección en función de la probabilidad de que sean relevantes a una necesidad de información de un usuario. La investigación en los modelos probabilísticos se ha centrado en utilizar las teorías formales de probabilidad y estadística para intentar estimar las probabilidades de relevancia.

Estos modelos resuelven el problema de RI a través del cálculo de probabilidades. La idea parte de la base de que para toda pregunta, la colección de documentos se divide en dos subconjuntos, el primero de ellos formado por todos los documentos que son relevantes a la pregunta y el segundo por todos aquellos documentos que no lo son. Así, el objetivo de un sistema de RI es localizar el primer subconjunto de documentos. La medida de similitud de un documento se define como la probabilidad que el documento se halle en el conjunto de documentos relevantes, dividido por la probabilidad de que se halle en el conjunto de documentos no relevantes. La idea se puede ver resumida en la figura 2.8.

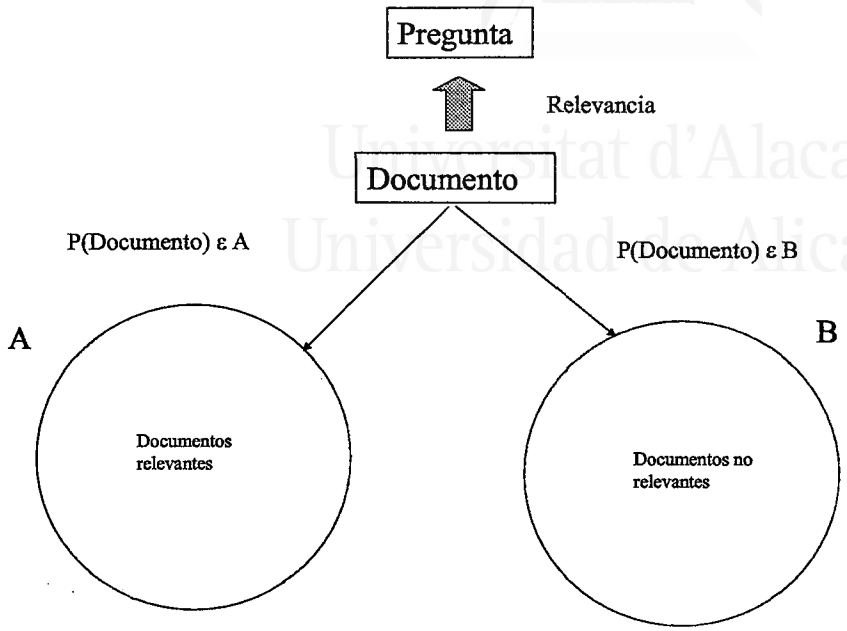


Figura 2.8. Modelo probabilístico

El objetivo de estos modelos es localizar aquellos documentos que maximicen la probabilidad de pertenecer al conjunto de documentos relevantes. No obstante, uno de los mayores obstáculos de los modelos probabilísticos es cómo estimar dichas probabilidades de forma computacionalmente eficiente.

En general, todos los modelos probabilísticos aplicados a la RI siguen el *principio de orden por probabilidades* (Roberston, 1977), el cual indica que el rendimiento óptimo de un sistema se consigue cuando los documentos son ordenados de acuerdo a las probabilidades de ser considerados relevantes. Esto inicialmente obliga a considerar la dependencias entre los diferentes documentos para poder realizar dicha ordenación. Normalmente, estas interdependencias se ignoran con el objeto de conseguir un sistema eficiente, de hecho los sistemas probabilísticos que las consideran no

consiguen mejorar los resultados, incrementando notablemente la complejidad del sistema (Cooper, 1991).

La medida más conocida, basada en una aproximación probabilística es la medida *okapi*. Esta ha sido utilizada con éxito en diferentes ediciones del TREC (S. Roberston y Beaulieu., 1998) y CLEF (Savoy, 2002). La definición completa de esta medida se halla en Roberston et al. (2000).

La medida de similitud del documento con la pregunta se obtiene de la siguiente forma

$$sim(q,d) = \sum_{j \in i \wedge q} (w_{i,j} \cdot w_{q,j}) \tag{2.12}$$

siendo

$w_{i,j}$ el peso asignado a un término t_j en un documento i.
$w_{q,j}$ el peso asignado a un término t_j en la pregunta q.

Ambas se calculan de la siguiente forma:

$$w_{i,j} = \frac{(k_1+1) \cdot f_{i,j}}{K + f_{i,j}} \tag{2.13}$$

donde K representa un valor proporcional entre la longitud documento y la media de la colección de documentos y se calcula de la siguiente forma:

$$K = k_1 \cdot ((1-b) + b \cdot \frac{l_i}{avdl}) \tag{2.14}$$

siendo

$f_{i,j}$ indica la frecuencia del término en el documento.
b y k_1 son constantes. En los experimentos realizados en la edición del CLEF del año 2001 en (Savoy, 2001) se fijan a 0,75 y 1,2 respectivamente.
D_i el tamaño del documento medido en bytes.
$advl$ la media del tamaño de la colección de documentos.

$$w_{q,j} = f_{q,j} \cdot \log_e(\frac{(N - df_j)}{df_j}) \tag{2.15}$$

donde

$f_{q,j}$ indica la frecuencia del término en la pregunta.
f_t es el número de documentos diferentes de la colección que contienen el término.
N es el número total de elementos de la colección.

2.3.4 Comparativa de medidas

Las medidas de similitud comentadas han sido evaluadas en diferentes ediciones de las conferencias TREC y CLEF y además se han realizado una serie de comparativas por investigadores ajenos a quienes las propusieron. En Kaszkiel y Zobel (2001) se utilizan diversas colecciones del TREC y se concluye que la medida más adecuada es la del coseno pivotado, aunque con la del sistema *okapi* se obtienen resultados similares. No obstante, en función de los experimentos realizados, en algunos el sistema que mejores resultados obtiene es el modelo del coseno pivotado (Voorhees y Harman, 1999), mientras que en otros es el modelo okapi (Savoy, 2002).

2.4 Técnicas de expansión de la pregunta

Para obtener la relevancia de un documento respecto a una pregunta, la mayoría de los modelos de RI se basan fundamentalmente en la aparición de los términos (o *stems* o *lemas*) de la pregunta en el propio documento. Como esta comparación se realiza a nivel de términos, uno de los mayores problemas que se tiene en la búsqueda de documentos relevantes a una pregunta es que frecuentemente los usuarios utilizan palabras diferentes en sus preguntas a las que utilizan los autores para describir los mismos conceptos en los documentos (Xu y Croft, 1996).

Por ello, uno de los objetivos a cumplir por los sistemas de RI es posibilitar la recuperación de documentos relevantes que no contienen exactamente los términos de la pregunta. Esto se consigue reformulando la pregunta o incorporando nuevos términos. Estas técnicas se denominan *expansión de la pregunta* (*query expansion*).

Bajo la expresión *expansión de preguntas* se engloban una serie de procesos automáticos o semiautomáticos que refinan la pregunta inicial, normalmente añadiendo una serie de términos adicionales, con el objetivo de mejorar el proceso de RI.

Las técnicas de expansión de la pregunta han permitido mejorar los resultados de los sistemas de RI. Los trabajos desarrollados en Harman (1988, 1992a) suponen un buen estudio comparativo de las diferentes técnicas de expansión existentes. Estas técnicas permiten recuperar documentos que no contienen exactamente las palabras que forman la pregunta, pero por otro lado, incurren en el riesgo de añadir términos a la pregunta expandida que no están muy relacionados con el objetivo original de la misma.

Existen diferentes técnicas de expansión de la pregunta. La diferencia fundamental entre ellas radica en la forma en la que se relacionan los términos que se añaden a la pregunta inicial y cómo se recalculan los pesos asignados a los términos originales y nuevos de la pregunta. Las técnicas más conocidas son las basadas en thesaurus, las de realimentación, las de análisis local y las de análisis global. Estas técnicas se comentan a continuación.

2.4.1 Basadas en thesaurus

El uso de thesaurus permite seleccionar palabras relacionadas con las que forman la pregunta a través de relaciones de sinonimia, hiponimia, etc.

Uno de los thesaurus más utilizados en este tipo de expansión es Wordnet (Miller et al., 1990; Miller, 1995). WordNet es una base de datos formada por relaciones semánticas entre las palabras, las cuales están agrupadas por sus significados. La relación central que utiliza WordNet para estructurar su información es la sinonimia. Wordnet ofrece el significado de las palabras mediante un conjunto de sinónimos que definen la palabra. No obstante, esta información no es suficiente y también se expresan otras relaciones entre las palabras como Mero/Holonimia, Hiper/Hiponimia, Antonimia, etc.

Los modelos basados en thesaurus no han logrado mejorar los resultados en tareas de RI, debido sobre todo a que pueden añadir

ruido a la pregunta, considerando relevantes, documentos que no lo son a la pregunta original, ya que cada palabra puede tener relaciones con diferentes términos en función del sentido que tenga en un contexto determinado. La mejora que pueden aportar estos sistemas es escasa, incluso aunque la búsqueda de relaciones entre palabras sea supervisada por humanos (Voorhees, 1994).

2.4.2 Realimentación

Los métodos de *realimentación* (*relevance feedback*) son la forma más popular de las estrategias de reformulación de las preguntas (Baeza-Yates y Ribeiro-Neto, 1999). Éstas se basan en mostrar al usuario los primeros documentos (entre 10 y 20) considerados relevantes por el sistema de RI empleando la pregunta original. El usuario especifica cuáles considera relevantes y cuáles no. Entonces se seleccionan aquellos términos que tienen un porcentaje de aparición alto en los documentos considerados relevantes y un porcentaje de aparición bajo en los no relevantes. Estos términos se añaden a la pregunta, la cual se vuelve a lanzar al sistema. Éste es un proceso repetitivo que permite mejorar sucesivamente la recuperación.

La principal ventaja consiste en limitar las funciones del usuario al hecho de indicar si un documento es relevante o no, evitando que tenga que ser él el que gestione la reformulación de la pregunta. Esta ventaja a su vez es su principal inconveniente, ya que obliga al usuario a intervenir en el proceso de refinamiento de la pregunta.

2.4.3 Análisis local

Las técnicas de análisis local (Jourlin et al., 1999; Xu y Croft, 1996; Chen, 2002) son similares a las de realimentación, pero en éstas no es necesaria la intervención del usuario. La idea es similar, es decir, seleccionar términos que aparecen en los documentos relevantes y no aparecen en los no relevantes. Sin embargo, el considerar qué documentos son relevantes, se efectúa de forma totalmente automática, ya que se consideran como relevantes los

primeros documentos recuperados al lanzar la pregunta original y como no relevantes el resto de documentos.

2.4.4 Análisis global

Mientras que las técnicas de análisis local sólo analizan una parte de la colección para expandir la pregunta (los primeros documentos relevantes recuperados), las técnicas de análisis global (Qiu y Frei, 1993) establecen relaciones de co-ocurrencia de todas las palabras en toda la colección. Estos modelos construyen en tiempo de indexación una matriz de co-ocurrencias en las que se almacenan las palabras que tienen un grado de aparición de forma conjunta en los documentos de la colección de forma elevada. Así cuando se lanza la pregunta se añaden a la misma aquellos términos que aparecen frecuentemente relacionados con las palabras de la pregunta en la colección de documentos.

2.4.5 Comparativa de las técnicas de expansión de la pregunta

Las técnicas de expansión de la pregunta comentadas permiten mejorar los resultados en la RI, facilitando la localización de documentos relevantes que no podrían ser localizados de forma directa y sí expandiendo la pregunta.

En (8) se puede ver un ejemplo de caso en el que se puede localizar un documento relevante utilizando un thesaurus.

(8) **Pregunta** : ¿Quién mató a Lincoln?
Documento : Lincoln fue asesinado mientras veía una obra de teatro por un conocido actor de teatro llamado John Wikes Booth

Accediendo al thesaurus con las palabras relevantes de la pregunta ("Lincoln", y "mató") podríamos obtener palabras que mantienen relaciones con esta última ("cargarse", "eliminar",

"asesinar", "sacrificar", etc.). Si estas palabras se añaden a la pregunta original, el documento citado en el ejemplo incrementaría notablemente sus probabilidades de ser considerado como relevante. No obstante, también hay que indicar que a su vez, también se incrementa la probabilidad de considerar como relevantes, documentos que no lo son.

Sin embargo, hay algunos casos en los que documentos relevantes pueden no ser recuperados utilizando este tipo de técnicas, ya que no aparecen relaciones semánticas entre los términos. En (9) se muestra un ejemplo de estos casos.

(9) **Pregunta**: La muerte del General Custer
Documento 1: El General Custer murió un domingo de Junio en Little Big Horn
Documento 2: El general Custer, también conocido como el general de los cabellos rubios fue el general más joven de la historia de los Estados Unidos
Documento 3 : El general cabellos rubios se autoinmoló, junto con sus hombres en las cercanías del río Little Big Horn

En dicho ejemplo se pueden ver tres documentos, de los cuales sólo el primero y el tercero son relevantes a la pregunta indicada, dado que cuando en este último se referencia a "Custer" se hace a través de uno de sus apodos "general cabellos rubios".

Si no se aplican técnicas de expansión, el tercer documento no sería considerado como relevante. No obstante, aplicando técnicas de análisis local o global o relevance feedback sí que se podrían establecer relaciones entre las palabras "Custer", "cabellos rubios", "Little Big Horn" que podrían, al expandir la pregunta original, considerar como relevante el tercer documento.

Estudios recientes (Buckley et al., 1995) han demostrado que la forma más eficaz de expandir automáticamente las preguntas son las de análisis local y global. Incluso hay modelos que combi-

nan ambas técnicas (Xu y Croft, 1996) con buenos resultados. La primera tiene la ventaja que trabaja con poca información adicional al sistema original (10 ó 20 documentos), mientras que la de análisis global requiere de unas estructuras de cierta complejidad espacial (matrices de co-ocurrencia). No obstante, los modelos de análisis global construyen estas estructuras en tiempo de indexación, afectando muy poco al tiempo de respuesta en la búsqueda, mientras que los sistemas de análisis local realizan el proceso adicional en tiempo de búsqueda, lo que puede incrementar el mismo.

2.5 ¿Pero qué desea el usuario de un sistema de RI?

En este capítulo se han revisado los principales objetivos y características de los sistemas de RI. No obstante, hay que considerar que los sistemas de RI van a ser utilizados por personas (usuarios) que pueden tener expectativas diferentes de las comentadas. En Croft (1995) se estudia el problema de la RI desde el punto de vista de lo que los usuarios esperan de un sistema de RI. En este artículo se indican los diez aspectos que los usuarios, en este caso los miembros de un centro de investigación en RI, consideran de mayor importancia en un sistema de RI. Estos aspectos ordenados de mayor a menor importancia son los siguientes:

1. Soluciones integradas
2. Recuperación de Información Distribuida
3. Eficiencia y Flexibilidad
4. "Magia"
5. Interfaz
6. Filtrado o Filtering
7. Eficacia
8. Recuperación de Información Multimedia
9. Extracción de Información
10. Realimentación

Las características más destacables de estos puntos son las siguientes:

Soluciones integradas. Un sistema de RI puede utilizarse para solucionar los problemas de almacenar y recuperar información. No obstante, esto puede distar de cubrir totalmente las necesidades de una empresa si no se utiliza junto con otras herramientas de manejo de información multimedia, sistemas OCR, Bases de Datos con información estructurada, etc. Uno de los objetivos de muchas empresas es integrar sus sistemas de RI con sus propios gestores de Bases de Datos y el resto de aplicaciones, ya que en caso contrario se pueden tener visiones parciales de toda la información de la que dispone.

Sistemas de Recuperación de Información distribuida. El concepto de información distribuida, supone que no toda la información se halla en un único punto (modelo centralizado). Esta circunstancia se puede producir en organizaciones en la que la información la almacena el grupo de trabajo que la genera. En consecuencia, uno de los problemas adicionales que el sistema de RI debe solucionar, es la selección de la mejor base de datos documental para localizar la información que se busca en cada momento.

Eficacia y eficiencia. Otros elementos valorados son la eficacia y eficiencia del sistema. Uno de los primeros problemas a determinar es cómo medir el rendimiento de los diversos sistemas de RI. Actualmente las dos medidas más referenciadas y utilizadas son las de cobertura y precisión, que serán ampliamente descritas en capítulo 5 de esta tesis. La obtención de los mejores resultados en estas medidas puede ser uno de los principales objetivos de los investigadores en RI, aunque no necesariamente, el de las empresas que desarrollan y venden dichos sistemas.

Por otro lado, los tiempos de proceso de indexación y búsqueda (eficiencia) tienen gran importancia a la hora de valorar un sistema de RI. Es de mayor importancia el tiempo de búsqueda que el de indexación, ya que este último puede realizarse *offline* (es decir sin interacción del usuario), mientras que el proceso de búsqueda suele requerir una respuesta *on-line* (mientras el usuario está esperando la respuesta). Además, dentro del concepto de eficiencia espacial, cabe citar la importancia de compactar o comprimir la

información almacenada para reducir la cantidad de espacio que debe ser utilizado para albergar la colección.

Curiosamente según el orden indicado, los usuarios valoran en mayor medida la idea de eficiencia que la de eficacia, lo cual indica que puede ser preferible una solución buena resuelta en tiempo real que una muy buena pero que tiene un coste temporal elevado. Actualmente es impensable que un buscador de Internet tenga éxito aun teniendo una eficacia notable, si obliga al usuario a esperar un tiempo considerable para obtener la respuesta.

La existencia de cada vez más trabajos (Witten et al., 1999; Kaszkiel et al., 1999; Moffat y Zobel, 1996; Zobel et al., 1995), que profundizan en la optimización de los procesos de indexación y búsqueda demuestra la importancia que tiene este tema.

Flexibilidad del sistema. Los usuarios también consideran importante la capacidad del sistema de RI para adaptarse a diferentes formatos de información ya sean multimedia o de texto. Este hecho se puede comprobar al acceder a la mayoría de buscadores en Internet, que no sólo se limitan a realizar búsquedas sobre documentos en formato texto, sino que lo amplían a documentos escritos en otros formatos tales como postscript o acrobat.

Extracción de información y filtrado. También cabe citar las referencias a otras técnicas de obtención automática de información, que son la extracción de información (ya comentada en el capítulo 1) y la de filtrado. Los sistemas de filtrado en vez de comparar preguntas con colecciones de documentos, comparan una serie de perfiles preestablecidos con documentos individuales, para determinar qué perfil se asigna a cada uno de ellos. Estos sistemas permiten clasificar los documentos en función de su contenido.

Realimentación y "magia". De los conceptos que el usuario considera importantes, cabe destacar los procesos que permiten recuperar documentos que no contengan exactamente los términos de las preguntas realizadas, esto suele ser denominado por los usuarios de RI como *magia*. Así, en la décima posición se marcan la realimentación y en el cuarto la "magia". Explícitamente, además de darle importancia a este hecho, se indica el orden de preferencia. Es decir, el usuario valora en mayor medida que este

proceso de expansión de pregunta se realice de forma automática, más que deba intervenir el mismo usuario en el proceso de refinamiento de la pregunta.

2.6 Conclusiones

En este capítulo se han revisado las principales características de los sistemas de RI. Dentro de esta revisión se han estudiado los principales modelos y medidas de similitud que incorporan los sistemas de RI más conocidos. También se han presentado las diversas técnicas que se suelen utilizar para mejorar los resultados obtenidos, así como la visión que los usuarios tienen de los sistemas de RI, indicando cuáles son los aspectos que en mayor medida valoran.

Los modelos de RI estudiados en este capítulo se basan en el análisis del documento completo para determinar la relevancia del mismo, es decir, valoran la aparición de los términos de la pregunta independientemente del lugar en el que aparezcan dentro del documento. Este análisis tiene algunos inconvenientes que serán comentados en el capítulo siguiente. También en este capítulo se estudiarán los modelos de RI basados en pasajes y la forma en la que estos modelos pueden subsanar algunos de estos inconvenientes.

3. Los sistemas de Recuperación de Información basados en pasajes

En este capítulo se realiza una introducción a los sistemas de recuperación de información basados en pasajes (RP). En primer lugar se estudian los inconvenientes que presentan los sistemas de RI que utilizan el documento completo como unidad de recuperación (RID). En segundo lugar se definen las bases sobre las que se asientan los sistemas RP para, posteriormente, clasificar y comentar las propuestas más importantes. Finalmente, se analizan las ventajas que aportan estos modelos frente a los de RID.

3.1 Inconvenientes del uso del documento completo como unidad de recuperación de información

Como ya se ha comentado en el capítulo anterior, para determinar el grado de relevancia de un documento con respecto a una pregunta, el principal aspecto que se valora es la cantidad de apariciones de los términos de la pregunta en el documento. Otros aspectos adicionales, que también se tienen en cuenta son, el peso o valor de cada término, calculado en función de su frecuencia de aparición en los documentos de la colección y, por último, factores de normalización que permiten comparar documentos de diferentes longitudes. Todos estos aspectos aplicados a la RID generan una serie de problemas:

- No tienen en cuenta la proximidad de aparición de las palabras de la pregunta en el documento.
- No localizan la zona realmente relevante del documento.

- Es difícil valorar la similitud de documentos de diferentes tamaños con respecto a la misma pregunta.

Estos inconvenientes se explican con mayor detalle en los siguientes subsecciones.

3.1.1 Proximidad de aparición de las palabras.

En líneas generales, los sistemas de RI valoran la aparición de las palabras de la pregunta en el documento, pero no tienen en cuenta el concepto de proximidad en la aparición de los términos de la pregunta en el mismo. El concepto de proximidad consiste en el hecho de que dichos términos aparezcan en posiciones muy cercanas dentro del documento.

Por ejemplo, dos documentos en los que aparece el mismo número de veces los términos de la pregunta, tendrán la misma valoración independientemente de si estos términos aparecen de forma próxima unos de otros o no. Un ejemplo citado en Salton y Allan (1994) indica que ante preguntas sobre John F. Kennedy (presidente de los Estados Unidos de América), era imposible eliminar los documentos recuperados sobre Anthony M. Kennedy (Juez de la corte suprema de justicia). Esto era debido a que al realizar un análisis global de los documentos era imposible discriminar qué documentos referenciaban a una persona o a otra. Ambos tenían el mismo apellido, fueron educados en la Universidad de Harvard y tenían alto grado dentro del gobierno de los Estados Unidos de América.

3.1.2 Localización de la parte relevante del documento.

Los sistemas RID determinan qué documento es relevante o no, pero no determinan la parte del documento que realmente es relevante a la pregunta. Un documento titulado "Biografía de Felipe II", con toda probabilidad será relevante con respecto a la pregunta "El nacimiento de Felipe II", pero sólo una parte del documento hará referencia concretamente a dicha pregunta. Además, como ya se ha comentado en el capítulo 1, cuando un documento es de tamaño considerable, el documento completo puede no ser una

unidad adecuada para visualizar la información requerida por el usuario.

3.1.3 Normalización del tamaño de los documentos.

No está suficientemente claro la forma óptima de valorar la similitud con respecto a una pregunta en los documentos de diferentes tamaños. De hecho algunos sistemas favorecen o discriminan a los documentos en función de su tamaño. Se ha demostrado que las posibilidades de recuperar un determinado documento utilizando una determinada medida de normalización, se desvían sistemáticamente de las probabilidades de relevancia a través de las diferentes colecciones. En Singhal et al. (1996) se demuestra, por ejemplo, que la medida del coseno tiende a favorecer la selección de documentos pequeños, ya que a medida que un documento crece en tamaño, los valores de normalización se elevan considerablemente. Como este factor se encuentra en el divisor de la medida de similitud, provoca la disminución de ésta.

3.2 Los sistemas de recuperación por pasajes

Una propuesta alternativa a los modelos RID para el cálculo de relevancia es considerar cada documento como un conjunto de pasajes, donde un pasaje se define como una porción o bloque contiguo de texto (Callan, 1994; Kaszkiel y Zobel, 2001). Estas propuestas determinan la relevancia de un documento respecto a una pregunta, estudiando la relevancia de cada uno de los pasajes que forman estos documentos. Así, un documento se divide en pasajes, de tal forma que cada pasaje es considerado como una entidad independiente cuando se valora su similitud, aunque evidentemente contenga enlaces al documento al que pertenece (Callan, 1994).

Los sistemas de RP pueden ser utilizados con dos objetivos diferentes. La primera aproximación consiste en determinar la relevancia de los documentos en función de la relevancia de los pasajes que los forman. Una segunda aproximación es la de suministrar respuestas concretas incluidas en un número pequeño de bytes,

a preguntas del estilo definido en los sistemas de BR (Kaszkiel et al., 1999).

3.2.1 Ventajas de los modelos de Recuperación de Pasajes.

A pesar que los sistemas de RP sólo consideren partes de cada documento para determinar su relevancia, aportan una serie de soluciones a los problemas detectados en los sistemas RID.

- Añade el concepto de proximidad al cálculo de relevancia.
- Define una nueva unidad de transmisión de información al más adecuada, tanto para un usuario como para un tratamiento posterior.
- Evitan los problemas de normalización de los documentos.

Valoración de la proximidad de la aparición de los términos de la pregunta. En general, los sistemas RID no contemplan el lugar del documento en el que los términos de la pregunta aparecen en el documento, sino que se limitan a contar el número de apariciones. Realmente, un aspecto que determina la relevancia de un documento con respecto a la pregunta es que sus términos aparezcan de forma cercana.

En la figura 3.1 se muestran dos documentos para los cuales se desea estudiar su relevancia con respecto a la pregunta "The death of General Custer". Ambos contienen las palabras relevantes de la pregunta ("death", "General" y "Custer"), pero en el segundo aparecen de forma próxima, con lo que es mucho más probable que la palabra "death" esté relacionada con el "General Custer" en el segundo documento que en el primero. No obstante, al estudiar de forma global el documento, la relevancia calculada para ambos documentos sería la misma.

Por el contrario, los sistemas de RP valoran los lugares en los que aparecen los términos de la pregunta dentro del documento, ya que al medir independientemente la similitud de cada pasaje con respecto a la pregunta, se tiene en cuenta que los términos que la forman aparezcan dentro de un mismo pasaje. Esto permite que en el ejemplo citado anteriormente se determine que el segundo

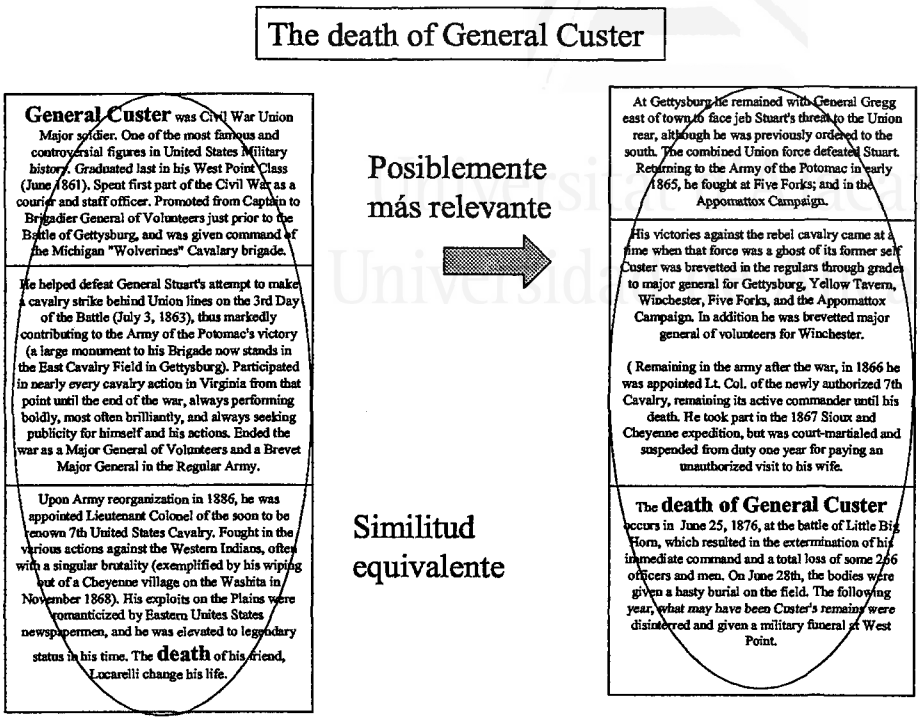


Figura 3.1. Valoración de la proximidad en sistemas de RI

documento sea más relevante que el primero, al aparecer todos los términos de la pregunta en un mismo pasaje (ver figura 3.2).

Definición de una unidad de transmisión más adecuada. La unidad que utilizan los sistemas RID es el documento completo. En consecuencia, indican el grado de relevancia de un documento, pero no facilitan información acerca de qué partes del mismo son las más relevantes.

Los sistemas de RP utilizan los pasajes como unidad de información. Esta unidad de transmisión, mucho más reducida, permite a un usuario concentrarse en los pasajes más relevantes y facilita la tarea a realizar por un sistema de BR que utilice dicha salida. Esto se puede ver en la figura 3.3, donde únicamente se suministra al sistema BR o al usuario el pasaje relevante.

Evitan problemas de normalización en el cálculo de similitud. Como se ha comentado, es difícil la comparación de valores

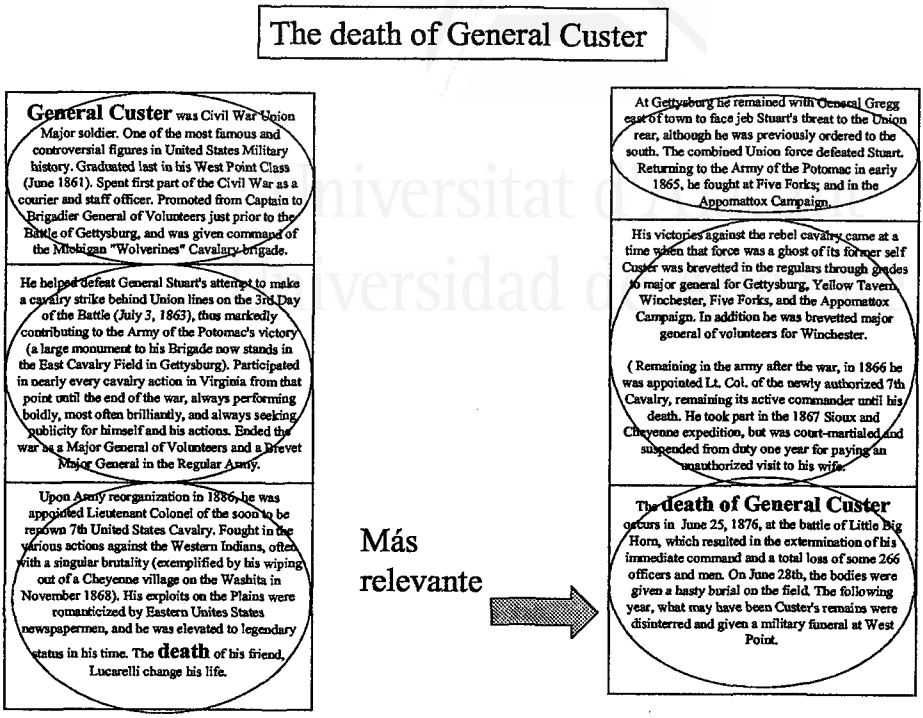


Figura 3.2. Valoración de la proximidad en los sistemas de RP

de similitud de documentos de diferentes tamaños. Sin embargo, los sistemas de RP pueden evitar en parte este problema, ya que las diferencias de tamaño entre los pasajes son sensiblemente menos significativas que las diferencias entre los documentos completos de una colección.

Entre otras cosas, esto puede provocar que un documento que tiene una pequeña parte muy relevante y el resto no lo es, puede ser considerado como no relevante por un sistema que se basa en el estudio completo del documento (ver figura 3.4).

En contraposición, esta circunstancia no suele producirse al emplear sistemas de RP puesto que valoran la similitud de los pasajes de forma independiente (ver figura 3.5).

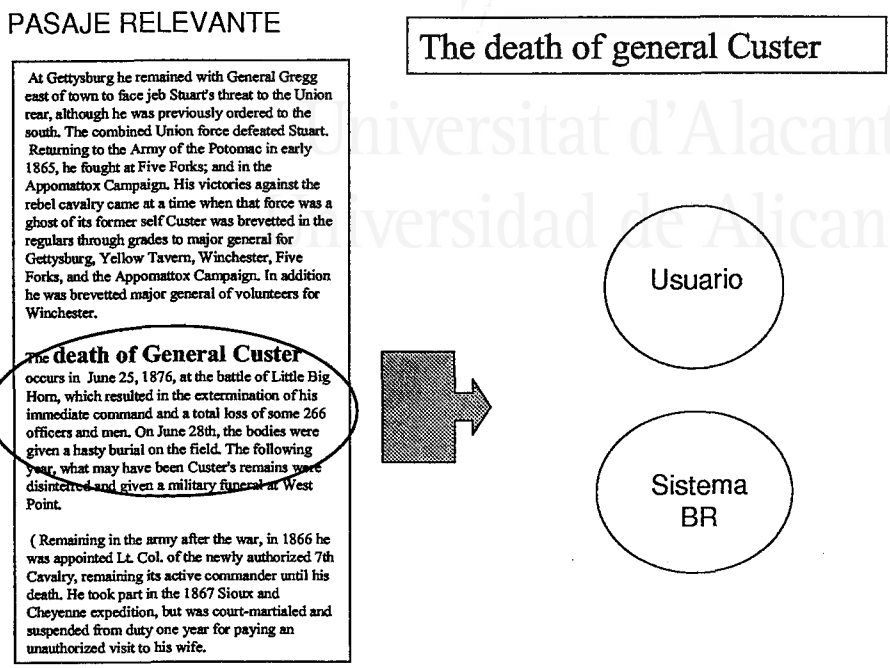


Figura 3.3. Unidad de transmisión de información sistemas de RP

3.3 Cálculo de similitud en los sistemas de RP

Los principales elementos que diferencian a los sistemas de RP se basan en dos aspectos fundamentales: la forma en la que cada uno de ellos define los pasajes en los que se divide el documento y en la forma de calcular la similitud de los documentos.

En la literatura podemos encontrar diferentes propuestas de sistemas de RP en función del procedimiento que se utiliza para definir los pasajes. No obstante, los modelos propuestos determinan la similitud del documento con respecto a un pregunta en función de la de los pasajes que lo forman siguiendo unas pautas comunes.

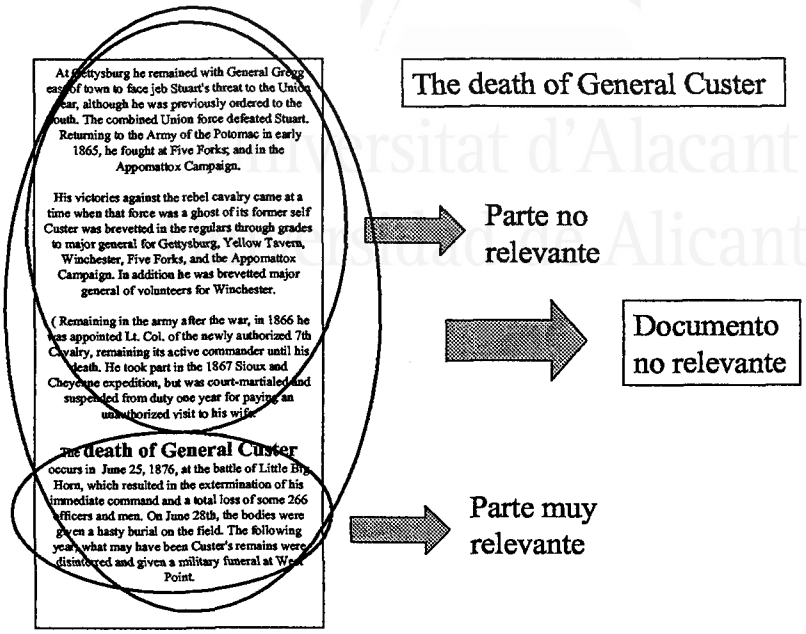


Figura 3.4. Valoración de fragmentos de texto relevantes en sistemas de RI

Los sistemas de RP definen, en primer lugar, los pasajes que forman el documento para luego valorar la similitud de cada uno de ellos con la pregunta. Dado que el objetivo principal es determinar la relevancia de un documento con respecto a una consulta, los sistemas de RP consideran relevante un documento siempre que contenga al menos un pasaje relevante. Además, la mayoría de los modelos propuestos, ordenan los documentos en función del mejor resultado obtenido por los pasajes que lo forman. No obstante, algunos modelos han experimentado con otras posibilidades.

Hearst y Plaunt (1993) probaron tanto seleccionar el valor de relevancia del mejor pasaje como el de la suma de las puntuaciones de todos los pasajes que forman un documento. Las conclusiones determinan que se obtienen mejores resultados determinando la relevancia de un documento en función de la mayor relevancia obtenida por alguno de sus pasajes.

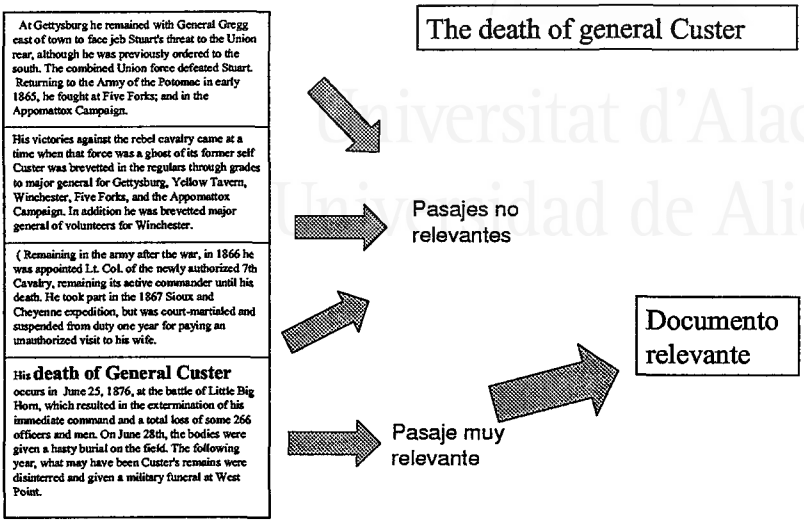


Figura 3.5. Valoración de fragmentos de texto relevantes en un sistema de RP

Callan (1994) propone un modelo en el que la similitud de un documento se obtiene calculando la similitud que obtiene el mismo utilizando un sistema RID y la del mejor pasaje utilizando un modelo RP. El valor total se obtiene asignando un peso de 1/3 al primer valor (documento) y un peso de 2/3 al segundo (mejor pasaje), permitiendo mejorar los resultados.

3.4 Clasificación de los sistemas basados en pasajes

En la literatura podemos encontrar diferentes modelos propuestos de RP en función de cómo se definen los pasajes en los que se divide el documento. Una clasificación generalmente aceptada es la definida en Callan (1994), que clasifica los sistemas de RP en tres modelos:

- **Modelos basados en el discurso.** Estos modelos utilizan las propiedades de estructura del documento para definir los pasajes.
- **Modelos semánticos.** Estos modelos definen los pasajes en función de las relaciones semánticas que se pueden establecer en los diferentes fragmentos del texto que forman el documento.
- **Modelos de ventana.** Estos modelos dividen los documentos en pasajes de tamaño fijo. En general los modelos de ventana intentan conseguir un conjunto homogéneo, en cuanto a tamaño, de los pasajes que se generan. Dentro de los modelos de ventana se hace una subclasificación adicional en Kaszkiel y Zobel (2001), donde se diferencian aquéllos que utilizan la estructura del documento en el momento de definir los pasajes o aquéllos que no la emplean.

3.4.1 Modelos basados en el discurso

Los modelos basados en el discurso utilizan elementos tales como frases, párrafos, secciones, marcas SGML y otros elementos del documento para definir los pasajes.

Salton et al. (1993) definen de dos formas los pasajes que forman un documento en base a las secciones o párrafos que lo forman. Así, en cada una de estas aproximaciones un documento se divide en tantos pasajes como secciones o párrafos tenga definidos. Utilizando este modelo sobre una serie de textos enciclopédicos se obtiene una mejora en la media de la precisión de un 9,7 y 18,5 por ciento aplicando la medida del coseno a secciones y párrafos respectivamente, con respecto al uso de esta misma medida sobre el documento completo.

Wilkinson (1994) realiza una división de los documentos en base a las marcas de sección que tienen algunos de los documentos de las colecciones del TREC. Posteriormente, aplica diferentes medidas de cálculo de relevancia, bien a las secciones, bien a los documentos completos. En estas pruebas se obtienen los mejores resultados cuando combina los valores de similitud obtenidos en base a los pasajes y documentos. No obstante, en el mismo artículo se comenta que la mejora introducida no es demasiado relevante.

Una aproximación similar, sería la definida en Brown y Yule (1983) en la que se definen los pasajes utilizando los encabezados y secciones que dividen un documento. No obstante, como se reconoce en este artículo, este tipo de aproximación tiene como principal inconveniente, el que no todos los textos disponen de una estructura de este tipo, limitando su ámbito de aplicación.

Callan (1994) realiza un estudio de los modelos del discurso utilizando los párrafos para la definición de los pasajes. La colección que se utiliza es una de las utilizadas en las conferencias TREC (colección TIPSTER). Esta colección contiene documentos de tamaño heterogéneo, siendo algunos de ellos bastante estructurados y otros no. Como en la colección de documentos no todos disponen de marcas de separación de párrafos, éstos se reconocen automáticamente en la etapa de indexación por medio de un conjunto de heurísticas, tales como el sangrado en primera línea. A los pasajes definidos se les calcula la similitud utilizando el sistema de RI "INQUERY" (Callan et al., 1992). Los resultados no mejoran los obtenidos por los modelos que se basan en el estudio documento completo.

Un ejemplo de la descomposición basada en el discurso se describe a continuación. En (10) se muestra un documento y en (11), (12) y (13) los pasajes en los que se descompondría si se utilizasen los párrafos como unidad de definición de los pasajes. La determinación de los párrafos se ha hecho en base a las líneas en blanco que aparecen en el documento.

(10) Custer, que estaba al frente del famoso Séptimo de Caballería, se ganó su sangrienta reputación en 1868, cuando fue enviado por el general Philip Sheridan -el "Oso Enfadado" de los cuarteles fronterizos- a sojuzgar a los indios de las praderas que se negaban a concentrarse en las reservas que el gobierno había establecido para ellos. Por qué se eligió a Custer para esta importante misión es un tema que se presta a conjeturas. Pues la carrera de

Custer como soldado había sido muy irregular.

Custer nació el 5 de diciembre de 1839 en New Rumley, Ohio. Se graduó en la Academia militar estadounidense de West Point, y gracias a la guerra civil -en la cual se distinguió en la persecución del general Robert E. Lee, comandante en Jefe de la Confederación-, alcanzó el grado de general de brigada a la temprana edad de 23 años.

A Custer se le subió el éxito a la cabeza. Se convirtió en un vanidoso, en un extravagante buscador de glorias. Se dejó crecer su rubia cabellera hasta los hombros y cubrió con sus propios retratos las paredes de su habitación. Cuando la guerra civil terminó, en 1865, el ego del general de brigada Custer se sintió gravemente herido, al ser rebajado al grado de capitán. Se convirtió en el hazmerreir de sus hombres, pero en el lapso de un año, había hecho méritos suficientes para recobrar el grado de teniente coronel.

(11) Custer, que estaba al frente del famoso Séptimo de Caballería, se ganó su sangrienta reputación en 1868, cuando fue enviado por el general Philip Sheridan -el "Oso Enfadado" de los cuarteles fronterizos- a sojuzgar a los indios de las praderas que se negaban a concentrarse en las reservas que el gobierno había establecido para ellos. Por qué se eligió a Custer para esta importante misión es un tema que se presta a conjeturas. Pues la carrera de Custer como soldado había sido muy irregular.

(12) Custer nació el 5 de diciembre de 1839 en New Rumley, Ohio. Se graduó en la Academia militar estadounidense de West Point, y gracias a la guerra civil -en la

cual se distinguió en la persecución del general Robert E. Lee, comandante en Jefe de la Confederación-, alcanzó el grado de general de brigada a la temprana edad de 23 años.

(13) A Custer se le subió el éxito a la cabeza. Se convirtió en un vanidoso, en un extravagante buscador de glorias. Se dejó crecer su rubia cabellera hasta los hombros y cubrió con sus propios retratos las paredes de su habitación. Cuando la guerra civil terminó, en 1865, el ego del general de brigada Custer se sintió gravemente herido, al ser rebajado al grado de capitán. Se convirtió en el hazmerreir de sus hombres, pero en el lapso de un año, había hecho méritos suficientes para recobrar el grado de teniente coronel.

Parece coherente pensar que los modelos del discurso van a ser efectivos en la definición de los pasajes ya que utilizan la propia estructura del documento. Sin embargo, se han detectado varios problemas con estos modelos.

El primer problema radica en el método definido para descomponer el documento en pasajes. Esta descomposición es sencilla siempre que la colección de documentos incorpore la suficiente información para realizar dicha descomposición. Es decir, si no se dispone de documentos que contienen las marcas que definen el inicio y final de los párrafos, realizar una descomposición en base a éstos será mucho más difícil e incluso puede llegar a ser casi imposible. Un ejemplo de este último caso son las colecciones de documentos basadas en boletines de noticias de radio (Ponte y Croft, 1997).

El segundo problema que se puede producir cuando se utilizan elementos del discurso para la división de pasajes es la gran variación de tamaños en los pasajes generados. En determinadas pruebas que utilizan colecciones muy heterogéneas se han genera-

do desde pasajes de una frase a pasajes formados por cientos de frases. Esta circunstancia elimina una de las ventajas que pueden aportar los sistemas de RP, que es la de homogeneizar los tamaños de los documentos a evaluar, ya que deben incorporar medidas de normalización en el cálculo de similitud de los pasajes en función de su tamaño.

Así, estos modelos no aseguran cierto grado de consistencia interna en los pasajes definidos (Salton y Allan, 1994). De hecho cuando se utilizan colecciones de documentos muy estructuradas se obtienen buenos resultados utilizando estos modelos (Salton et al., 1993). Pero cuando se utilizan documentos en los que la estructura no garantiza esa consistencia de los pasajes, por ejemplo cuando los párrafos se utilizan más por motivos visuales que por contenido, los resultados que se obtienen son sensiblemente inferiores (Kaszkiel y Zobel, 2001).

3.4.2 Modelos semánticos

El objetivo de los modelos semánticos es conseguir que los pasajes en los que se divide el documento tengan un alto grado de consistencia interna en su contenido. Para ello, estos modelos se basan en definir los pasajes en función de las relaciones semánticas que se pueden establecer en los diferentes fragmentos del texto.

Generalmente, para conseguir esa consistencia interna, estos modelos estudian las relaciones entre párrafos adyacentes y establecen si existen relaciones semánticas entre ellos. En caso afirmativo los párrafos se unen formando un único pasaje.

En Hearst y Plaunt (1993) y Hearst (1994) se definen una serie de algoritmos que permiten definir los pasajes formados por una serie de párrafos consecutivos con relaciones semánticas entre ellos. A cada uno de estos pasajes se les da el nombre de *tiles*. Básicamente el método aplica un cálculo de similitud, utilizando la medida del coseno, para determinar si dos párrafos consecutivos pertenecen al mismo *tile*. Este modelo rompe el documento inicialmente en bloques de un determinado tamaño, para después calcular su similitud utilizando la fórmula del coseno. Luego suaviza los resultados obtenidos para determinar los potenciales límites de

cada pasaje. Este algoritmo denominado *TextTiling* de generación de pasajes es muy conocido y existen diversas implementaciones del mismo[1]. K. Richmond y Amitay (1997) proponen una ligera modificación a dicho algoritmo, otorgando un peso a cada palabra en función de la aparición de las palabras en el documento.

Salton et al. (1996) introduce el concepto de *tema* para definir los pasajes. Este modelo no sólo se limita a calcular las similitudes utilizando también la medida del coseno entre párrafos consecutivos, sino que las realiza entre todos los párrafos del documento para detectar relaciones entre todos ellos. De esta forma, se definen los *temas* como un conjunto de párrafos, no necesariamente adyacentes, relacionados semánticamente entre sí. No obstante, dado que estos cálculos serían muy elevados en caso de documentos grandes, se limita el cálculo de relaciones a los cinco párrafos adyacentes. Este modelo obtiene mejores resultados en preguntas con cierta complejidad que en la recuperación del documento completo. No obstante, en preguntas sencillas los resultados son sensiblemente peores.

Así, en general, los modelos semánticos son capaces de determinar relaciones entre más de un párrafo al estudiar si ambos comparten ciertas palabras. Supongamos el documento descrito en (14).

> (14) Allí, en la cima de ese promontorio -que ahora se llama colina Custer- apareció Caballo Loco con 1.000 guerreros a caballo. Por un momento, los indígenas contemplaron con desdén a Custer y a la banda dispersa en que se había convertido su exhausta caballería. Luego, dando feroces alaridos, los indígenas cargaron colina abajo.
>
> La caballería de Custer fue reducida en pocos segundos. Sus soldados desmontaron e intentaron defenderse en campo abierto, sin apenas protección. Lucharon con valentía, tratando de conservar sus caballos. Pero a medida que la gritería de los Sioux se acercaba, los jinetes

[1] http://www.cs.man.ac.uk/~mary/choif/software.html

de Custer tuvieron que liberar las cabalgaduras. Ahora no existía esperanza de escapar.

Los orgullosos soldados de caballería quedaron reducidos a un puñado. En los aledaños de la batalla, algunos pocos soldados heridos levantaron sus brazos y pidieron ser tomados prisioneros. Pero no hubo prisioneros ese día. Los heridos fueron muertos a tiros o a cuchilladas. El general fue uno de los últimos en morir. A medida que mermaban sus filas y los indígenas se le acercaban, vieron que Pahuska ya no tenía el cabello largo hasta los hombros. Se lo había cortado, y esa era la razón por la cual los atacantes no lo habían reconocido de inmediato.

Dicho documento está formado por tres párrafos, los modelos semánticos intentarían comprobar si existe alguna relación entre los mismos. Así, tanto en el primero como en el segundo párrafo aparece varias veces (*tema*) la palabra "Custer". Esto permitiría determinar que ambos párrafos formarán parte de un mismo pasaje. No así, el tercer párrafo que inicialmente parece no tratar el mismo tema. De esta forma dicho documento se dividiría en los pasajes descritos en (15) y (16).

(15) Allí, en la cima de ese promontorio -que ahora se llama colina Custer- apareció Caballo Loco con 1.000 guerreros a caballo. Por un momento, los indígenas contemplaron con desdén a Custer y a la banda dispersa en que se había convertido su exhausta caballería. Luego, dando feroces alaridos, los indígenas cargaron colina abajo.

La caballería de Custer fue reducida en pocos segundos. Los soldados desmontaron e intentaron defenderse en campo abierto, sin apenas protección. Lucharon con valentía, tratando de conservar sus caballos. Pero a medida que la gritería de los Sioux se acercaba, los jinetes de Custer tuvieron que liberar las cabalgaduras. Ahora

no existía esperanza de escapar. Estos métodos aplicados al ejemplo siguiente, podrían determinar que algunos párrafos pertenecieran al mismo segmento o pasaje, ya que comparten (entre otras) la palabra Custer.

(16) Los orgullosos soldados de caballería quedaron reducidos a un puñado. En los aledaños de la batalla, algunos pocos soldados heridos levantaron sus brazos y pidieron ser tomados prisioneros. Pero no hubo prisioneros ese día. Los heridos fueron muertos a tiros o a cuchilladas. El general fue uno de los últimos en morir. A medida que mermaban sus filas y los indígenas se le acercaban, vieron que Pahuska ya no tenía el cabello largo hasta los hombros. Se lo había cortado, y esa era la razón por la cual los atacantes no lo habían reconocido de inmediato.

No obstante, puede ocurrir que diversos párrafos compartan el mismo *tema* aunque no compartan palabras, como ocurre en el ejemplo anterior. Al observar el segundo pasaje, se puede comprobar que tiene una alta relación con el primero. En este segundo pasaje se habla del General (obviamente Custer) y de Pahuska (un apodo que le dieron los indios a Custer). No obstante al no contener palabras similares, se consideran semánticamente disjuntos.

Para solucionar este problema, Ponte y Croft (1997) propone utilizar técnicas de expansión de la pregunta, en concreto las técnicas de *análisis local*, propuestas por Xu y Croft (1996). Estas técnicas buscan palabras relacionadas para los términos contenidos en un párrafo (en este caso una ventana de tamaño fija). Estas palabras relacionadas se obtienen en base a las co-ocurrencias de las mismas en una serie de colecciones. Para determinar si dos párrafos pertenecen al mismo tema se aplican cálculos de similitud entre ellos y las palabras relacionadas obtenidas de cada

párrafo. Este método soluciona los problemas citados, cuando los diversos párrafos tienen pocas palabras en común, obteniendo mejoras en precisión y cobertura de hasta un 7% sobre el algoritmo *TextTiling* original.

Si bien, los modelos semánticos pueden solucionar el hecho que no se disponga de la estructura de los documentos a analizar, sus mayores inconvenientes son, por una parte que incrementan la complejidad del proceso de indexación y, por otra, su todavía baja eficacia.

El incremento en la complejidad se debe a la necesidad de aplicar algoritmos adicionales a cada documento para determinar las relaciones entre los diferentes párrafos. No obstante, es un inconveniente menor ya que este proceso se realiza una única vez, antes de la indexación.

El segundo inconveniente es que todavía los algoritmos de segmentación basados en información semántica están muy lejos de ser seguros (Kaszkiel y Zobel, 2001). Incluso algunos de los autores de los algoritmos más conocidos indican que esta segmentación automática todavía tiene unos resultados bastante alejados en cuanto a exactitud de lo que sería una segmentación realizada de forma manual (Hearst, 1994).

3.4.3 Modelos de ventana

Los modelos de ventana dividen los documentos en un conjunto de pasajes de tamaño homogéneo. Para ello, se define un rango de tamaño que no sea ni demasiado pequeño ni demasiado grande, y se procura que todos los pasajes tengan un tamaño comprendido entre los límites fijados por dicho intervalo. Con ello, se evitan los problemas de utilizar medidas de normalización del tamaño de los pasajes.

Dentro de los modelos de ventana Kaszkiel y Zobel (2001) hacen una subclasificación adicional en la que se diferencian aquéllos que utilizan la estructura del documento en el momento de definir los pasajes o aquéllos que no la utilizan.

Modelos de ventana basados en la estructura del documento. Dentro de estos modelos cabe citar los trabajos de Callan

(1994), Zobel et al. (1995) y Moffat et al. (1993), que utilizan los párrafos como unidad para definir los pasajes. En dichos modelos se define un tamaño, de tal forma que los párrafos que lo superan forman un único pasaje y los párrafos de tamaño menor se unen con párrafos consecutivos para formar un pasaje.

Callan (1994) define el modelo *bounded paragraphs*. Este modelo define el tamaño del pasaje en base al número de palabras que lo forman. Experimentalmente se demuestra que utilizando pasajes con tamaños comprendidos entre 50 y 200 palabras se obtienen los mejores resultados. Así, cualquier párrafo que tenga un tamaño inferior se une con el siguiente.

Los modelos desarrollados por (Moffat et al., 1993) y (Zobel et al., 1995) son similares al anterior, pero utilizan para la definición de los pasajes el número de bytes que lo forman como unidad umbral. El tamaño de los pasajes puede quedar, según ambas aproximaciones sobre 1.000 ó 2.000 bytes respectivamente.

Estas propuestas tienen como principal objetivo garantizar que el tamaño de todos los pasajes sea similar. Así, dado el documento descrito en (17). Si se considerara el tamaño de los pasajes de 50 palabras, la división de este documento se realizaría generando los pasajes (18), (19) y (20). Esta división se realiza de la siguiente forma, el primer párrafo ya tiene un tamaño superior al umbral fijado, no así el segundo y el tercer párrafo que en cuyo caso se unirán formando un único pasaje.

(17) Custer fue uno de los últimos en morir. A medida que mermaban sus filas y los indígenas se le acercaban, vieron que Pahuska ya no tenía el cabello largo hasta los hombros. Se lo había cortado, y esa era la razón por la cual los atacantes no lo habían reconocido de inmediato. El general estaba en el centro de un pequeño, patético grupo de soldados sobrevivientes.

Toro Sentado comentó luego: "Donde se cumplió la última batalla, el de los largos cabellos estaba como una gavilla de trigo con todas las espigas despenachadas a

su alrededor". Muy pronto, Custer fue cubierto por una oleada de guerreros indígenas.
Muchos indios reclamaban más tarde haber sido quienes dieron muerte al odiado Pahuska. Era un legítimo motivo de orgullo.
Caballo Loco se trasladó a una reserva y se sometió a los blancos. Pero fue arrestado y luego asesinado a bayonetazos mientras trataba de escapar del Fuerte Robinson, en 1887. Sus últimas palabras fueron: "Dejadme ir, amigos míos. Ya me habéis hecho suficiente daño". Toro Sentado huyó con 3.000 guerreros al Canadá, la "Tierra de la Gran Madrina", la reina Victoria.

(18) Custer fue uno de los últimos en morir. A medida que mermaban sus filas y los indígenas se le acercaban, vieron que Pahuska ya no tenía el cabello largo hasta los hombros. Se lo había cortado, y esa era la razón por la cual los atacantes no lo habían reconocido de inmediato. El general estaba en el centro de un pequeño, patético grupo de soldados sobrevivientes.

(19) Toro Sentado comentó luego: "Donde se cumplió la última batalla, el de los largos cabellos estaba como una gavilla de trigo con todas las espigas despenachadas a su alrededor". Muy pronto, Custer fue cubierto por una oleada de guerreros indígenas.

Muchos indios reclamaban más tarde haber sido quienes dieron muerte al odiado Pahuska. Era un legítimo motivo de orgullo.

(20) Caballo Loco se trasladó a una reserva y se sometió a los blancos. Pero fue arrestado y luego asesinado a bayonetazos mientras trataba de escapar del Fuerte Robinson, en 1887. Sus últimas palabras fueron: "Dejadme ir, amigos míos. Ya me habéis hecho suficiente daño". Toro Sentado huyó con 3.000 guerreros al Canadá, la "Tierra de la Gran Madrina", la reina Victoria.

De esta forma se consigue que todos los pasajes tengan un tamaño similar, con lo que los posibles efectos perniciosos de la normalización de documentos son menores.

Modelos de ventana no basados en la estructura del documento. En general, un inconveniente de los modelos basados en la estructura del documento es que se debe conocer dicha estructura. Dentro de los modelos de ventana existen una serie de propuestas que definen los pasajes sin considerar esta estructura. Los más conocidos son los denominadas *sliding windows* (Callan, 1994) y *arbitrary passages* (Kaszkiel y Zobel, 1997, 2001).

El modelo *sliding windows* define el primer pasaje del documento de tal forma que empieza en la primera palabra del documento que coincide con una de las palabras de la pregunta. A partir de dicha definición se definen pasajes de un tamaño fijo de palabras (denominado n). Cada pasaje se solapa con el anterior (la idea de solapamiento de pasajes se ampliará en este mismo apartado). Los sucesivos pasajes empiezan cada n/2 palabras, es decir el segundo pasaje empieza en la mitad del anterior y finaliza n palabras después. Experimentalmente se comprueba que el tamaño del pasaje con el que se obtienen mejores resultados depende de la colección de documentos que se está utilizando, encontrándose este valor entre cincuenta y trescientas palabras. No obstante, se indica que utilizando una única definición de pasaje o ventana de doscientas o doscientas cincuenta palabras se obtienen buenos resultados. Las mejoras obtenidas llegan a ser de hasta un 20% con respecto a los modelos RID.

El modelo *arbitrary passages* también se basa en la definición de ventanas que están formadas por un número determinado de

palabras. En esta propuesta se define un *arbitrary passage* o *pasaje arbitrario* como una secuencia, de cualquier longitud, de palabras que empiezan en cualquier palabra del documento. Posteriormente se definen dos modelos de pasajes arbitrarios los de longitud fija y los de longitud variable.

El modelo de *longitud fija* (*fixed length arbitrary passages*) es un modelo muy similar al de *sliding windows* citado previamente. Por otra parte, el modelo de *longitud variable* es una aproximación más flexible, en la que se utilizan pasajes de diferente tamaño y se selecciona el mejor para representar a un documento. Para ello se definen diferentes tamaños de pasajes, en los que se descompone cada documento. La similitud de estos pasajes se calcula utilizando *la medida del coseno pivotado*. Es decir, si se utilizan pasajes de 25, 50 y 100 palabras, cada documento se divide en pasajes de dichos tamaños y se calcula la similitud de cada uno de ellos. Posteriormente a cada documento se le asigna la similitud del pasaje que obtiene mejores resultados. Hay que destacar que en este modelo sí que es necesario utilizar factores de normalización, ya que se comparan pasajes de tamaños sensiblemente diferentes. Este modelo obtiene una ligera mejoría en los resultados con respecto al *fixed length arbitrary passages* entre un 1% y 5%, pero sólo en preguntas cortas.

Otro aspecto que caracteriza los modelos de ventana es el de permitir el solapamiento. En general, los modelos de pasajes tienen el inconveniente de que una segmentación inadecuada puede provocar que documentos relevantes queden mal clasificados, debido a que las palabras de la pregunta se hallen en pasajes contiguos (ver figura 3.6). En dicha figura se puede observar, que las palabras de la pregunta se hallan muy cerca, pero en pasajes diferentes, lo cual puede provocar una medida de similitud baja para un documento relevante.

Para solucionar este problema, aparece la idea de solapamiento de pasajes. Esta idea implica que diversos fragmentos del documento se repiten en más de un pasaje consecutivo. Así, se puede evitar que apariciones cercanas de las palabras de la pregunta no se encuentren de forma completa en ningún pasaje. La comparativa de lo que supondría utilizar un esquema sin y con solapamiento

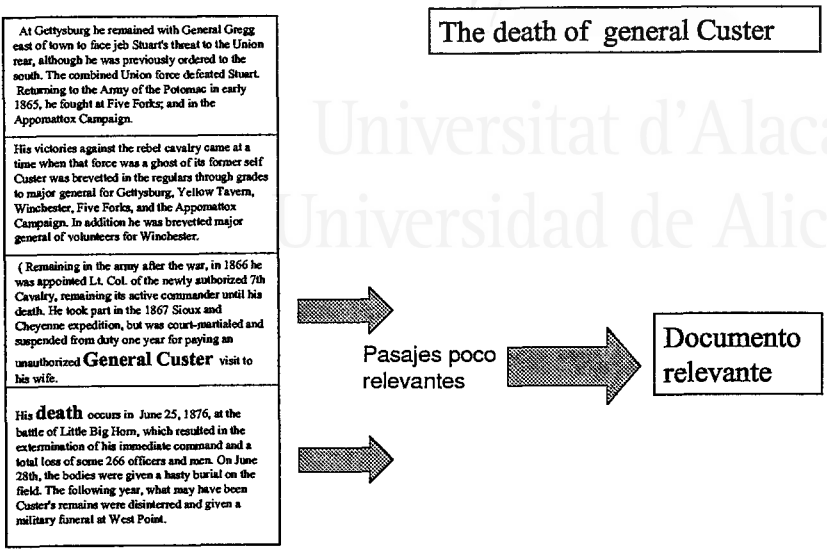


Figura 3.6. Problemática de los sistemas de RP

se puede ver en la figura 3.7. En este gráfico el mismo documento se convierte en el primer esquema en cuatro pasajes que no comparten ningún fragmento del documento, mientras que en el segundo se generan una serie de pasajes que comparten contenido del documento. Como se puede observar, se define un pasaje, marcado con una elipse de mayor grosor, muy relevante a la pregunta "The death of General Custer".

Esta circunstancia, evidentemente, incrementa los tiempos de cómputo y la cantidad de almacenamiento necesaria al incrementarse notablemente el número de pasajes a evaluar. Para solucionar esto se han propuesto diversas técnicas (Moffat y Zobel, 1996). Algunas de estas técnicas no almacenan realmente todos los pasajes generados sino que realizan esta descomposición en tiempo de ejecución. Esto no sólo permite disminuir la cantidad de almacenamiento necesaria, de hecho será similar a la que se utilizarían en caso de análisis del documento completo, sino que además permite definir el pasaje en tiempo de ejecución, esto fa-

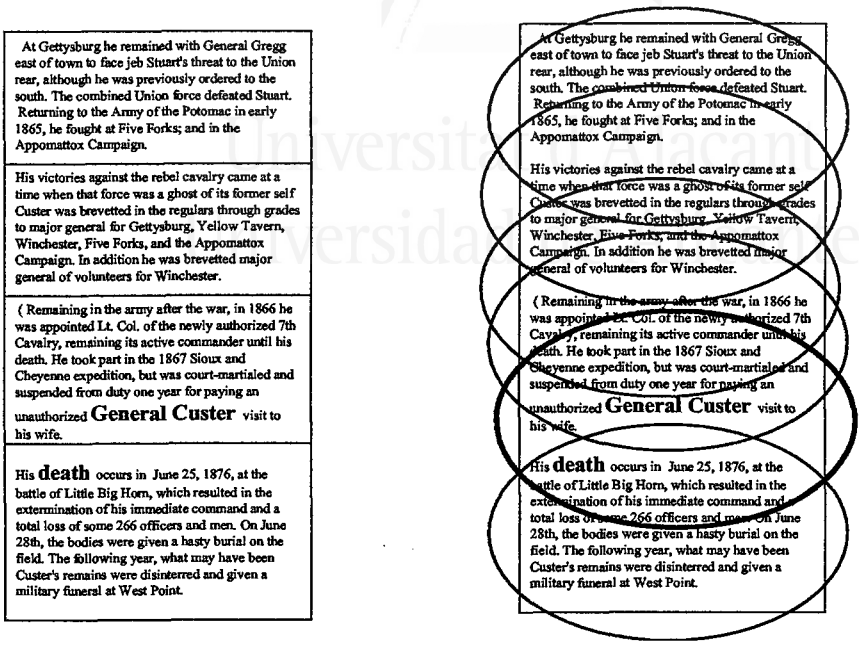


Figura 3.7. Modelos sin y con solapamiento

cilita la adecuación del tamaño del pasaje al tamaño de pregunta que se realiza.

Dentro de los modelos que utilizan pasajes solapados cabe citar los de *sliding windows* (Callan, 1994) y *arbitrary passages* (Kaszkiel y Zobel, 2001), ambos ya comentados en el apartado anterior.

Estos modelos deben definir el grado de solapamiento, es decir a partir de qué parte de un pasaje su contenido también formará parte del siguiente pasaje. Como ya se ha comentado, el modelo de *sliding windows* se basa en definir que el primer pasaje del documento empieza en la primera palabra del documento que coincide con una de las palabras de la pregunta. A partir de dicha definición se definen pasajes de tamaño n palabras, que empiezan cada n/2 palabras. Un ejemplo de utilización de este modelo se puede ver a continuación (suponiendo que los pasajes estuviesen formados por 50 palabras), descomponiendo el mismo documen-

to ejemplo descrito en (17). Así, los primeros pasajes serían los descritos en (21), (22) y (23).

(21) Custer fue uno de los últimos en morir. A medida que mermaban sus filas y los indígenas se le acercaban, vieron que Pahuska ya no tenía el cabello largo hasta los hombros. Se lo había cortado, y esa era la razón por la cual los atacantes no lo habían reconocido

(22) tenía el cabello largo hasta los hombros. Se lo había cortado, y esa era la razón por la cual los atacantes no lo habían reconocido de inmediato. El general estaba en el centro de un pequeño, patético grupo de soldados sobrevivientes.

Toro Sentado comentó luego: "Donde se cumplió la última batalla,

(23) de inmediato. El general estaba en el centro de un pequeño, patético grupo de soldados sobrevivientes. Toro Sentado comentó luego: "Donde se cumplió la última batalla,el de los largos cabellos estaba como una gavilla de trigo con todas las espigas despenachadas a su alrededor". Muy pronto, Custer fue cubierto por

En este ejemplo se puede observar cómo cada pasaje comparte un número determinado de palabras con el anterior.

Los modelos de ventana son, como contrapunto a los semánticos o del discurso, mucho más fáciles de construir (Kaszkiel y Zobel, 1997) ya que pueden ser aplicados de forma directa a documentos de los que no se disponga de información sobre su estructura, o incluso, que carezcan de ella.

Dentro de los modelos de ventana, en general obtienen mejores resultados aquéllos que las utilizan de forma solapada, aunque éstos como contrapunto, incrementan la complejidad del modelo. Las propuestas realizadas de estos modelos suelen construir los pasajes sin atender a la estructura del documento. Por ello, el problema que pueden tener es la pérdida de parte de información del contexto. Así, si los pasajes que se obtienen se muestran directamente al usuario o son utilizados como entrada por un sistema de BR pueden no ser totalmente comprendidos.

Modelo	Sistema	Unidad	Método
Discurso	Párrafos	Párrafos	Los pasajes se forman por un párrafo del documento.
	Secciones	Marcas HTML	Los pasajes se forman en base a las secciones definidas en el documento.
Semántico	TextTiling	Párrafo	Los pasajes se forman por la unión de párrafos consecutivos que traten el mismo tópico.
Ventana	Pages	Párrafo + Caracteres	Los pasajes se forman con la unión de párrafos consecutivos de forma que no superen un tamaño determinado medido en caracteres.
	Bounded Paragraphs	Párrafo + Palabras	Los pasajes se forman con la unión de párrafos consecutivos de forma que no superen un tamaño determinado medido en palabras.
	Sliding Windows	Palabras	Los pasajes se forman con un número fijo de palabras consecutivas.
	Arbitrary Passages	Palabras	Los pasajes se forman con un número variable de palabras consecutivas
	IR-n	Frases	Los pasajes se forman por un número determinado de frases consecutivas.

Tabla 3.1. Definición de pasajes en modelos de RP más representativos

En la tabla 3.1 se indican de forma esquemática la forma utilizada para definir los pasajes por los modelos de RP más representativos. Como se puede ver, fundamentalmente se utilizan como unidad los párrafos, a veces apoyados con el uso de otros elementos como las palabra y caracteres para garantizar un tamaño homogéneo en los pasajes definidos. Los modelos de ventana que

no se basan en elementos del discurso utilizan un número de palabras para definir los pasajes.

3.5 Conclusiones

En este capítulo se han revisado las principales ventajas que suponen la utilización de los modelos de RI basados en pasajes frente a los modelos que se basan en el documento completo. Por los experimentos realizados en la mayoría de las propuestas de sistemas basados en pasajes, se concluye que la aplicación de éstos suele mejorar el rendimiento de los sistemas basados en el documento completo. Estas mejoras se acentúan sobre todo en colecciones de documentos poco estructuradas y/o cuyos tamaños son heterogéneos, lo que demuestra los problemas de normalización en los cálculos de relevancia en documentos de diferentes tamaños.

También cabe destacar, que además de las ventajas en cuanto a la mejora de la eficacia del sistema, los sistemas de RP ofrecen una mejor unidad de transmisión (pasaje en vez de documento) para su posterior aplicación como entrada para sistemas de BR o para mostrar a un usuario.

Sin embargo, los modelos de RP tienen una serie de inconvenientes, entre los que destacan el incremento de la complejidad de los procesos de indexación y búsqueda. En los primeros ya que añaden el proceso de segmentación de los documentos, y en los segundos al incrementar el número de elementos (pasajes) para los que hay que realizar los cálculos de relevancia. Generalmente, estos inconvenientes se recompensan con mejores resultados en la recuperación.

Otro aspecto tratado en el tema ha sido la clasificación y diferenciación de los diferentes modelos de RP propuestos. Como se ha comentado cada modelo tiene sus ventajas e inconvenientes, teniendo mayor o menor éxito su aplicación en función de los tipos de colecciones.

Finalmente, habría que destacar las recientes investigaciones en los modelos de ventana que utilizan ventanas solapadas, los cuales consiguen una mejoría notable tanto en resultados como

Volver al índice/Tornar a l'índex

en flexibilidad del sistema, a pesar del incremento de complejidad que suponen.

4. IR-n: Definición del sistema

En este capítulo se presenta el sistema IR-n que es el fruto de la investigación realizada en esta tesis. Este sistema se encuadra dentro de los sistemas de RI basados en pasajes. La principal diferencia con otros sistemas pertenecientes a la misma categoría radica en el método propuesto para la definición de los pasajes y para el cálculo de similitud entre los documentos y las preguntas.

La unidad que utiliza el sistema IR-n para definir los pasajes es la frase. Así, los pasajes se definen mediante un número determinado de frases consecutivas del documento. Este modelo permite recuperar pasajes con entidad sintáctica propia, fácilmente entendibles por un usuario y tratables por un sistema de BR. Además, es un modelo muy flexible y adaptable a diferentes colecciones de documentos y tipos de preguntas. Por otra parte, la medida de similitud empleada permite aplicar conceptos de proximidad de aparición de los términos de la pregunta en el documento.

Este capítulo se ha dividido en cinco partes. La primera de ellas realiza un repaso a las características fundamentales de los modelos de RID y RP. En la segunda se define el modelo conceptual del sistema IR-n, es decir, el conjunto de características básicas del mismo. La tercera parte profundiza en los aspectos de implementación del sistema. La cuarta parte define una serie de características que se han añadido al sistema IR-n con el objetivo de mejorar su rendimiento. Finalmente, la última parte muestra las conclusiones del capítulo.

4.1 Introducción

Los sistemas RID y RP, tal y como se ha descrito en los capítulos 2 y 3 respectivamente, contemplan dos enfoques diferentes en la forma de abordar el problema de la localización de documentos relevantes ante una pregunta o necesidad de información del usuario. En general, el concepto de relevancia se sustenta en una cuantificación de la misma. De forma esquemática, dados un documento D y una pregunta Q, el objetivo final es medir la similitud o relevancia entre ambos:

$$sim(D, Q) =? \tag{4.1}$$

Para determinar dicha relevancia, los sistemas RID aplican directamente una serie de funciones, denominadas medidas de similitud, que cuantifican ese valor de relevancia entre el documento y la pregunta. Fundamentalmente, estas medidas se basan en la cantidad de términos que comparten el documento y la pregunta.

Para realizar la misma tarea, los sistemas de RP realizan una serie de pasos adicionales:

1. Descomponer el documento D en pasajes

   $$D \rightarrow P_1..P_n \tag{4.2}$$

   siendo P_i un fragmento contiguo de texto del documento D.
2. Calcular la similitud X_i de todos los pasajes con respecto a la pregunta

   $$\forall_{i \in 1..n} \rightarrow sim(P_i, Q) = X_i \tag{4.3}$$

3. Calcular la similitud del documento en función de la similitud de todos los pasajes que forman el documento

   $$sim(D, Q) = f(X_1..X_n) \tag{4.4}$$

Como se puede comprobar, los modelos de RP tienen una complejidad mayor que la de los sistemas RID dado que deben realizar una serie de tareas adicionales. No obstante, los modelos de RP ofrecen una serie de ventajas (descritas en el capítulo 3), que compensan el incremento de complejidad con un incremento del rendimiento obtenido.

Las tareas que realizan los sistemas de RP para obtener la similitud de un documento con respecto a una pregunta, se han abordado desde diferentes ópticas. Fundamentalmente, el aspecto que más distingue las distintas propuestas radica en el concepto de pasaje que proponen, aunque también existen diferencias en las otras dos tareas.

El modelo de RP desarrollado en esta tesis realiza una serie de nuevas propuestas para la realización de cada una de las tareas, que se describirán en las siguientes secciones.

4.2 El sistema de recuperación por pasajes IR-n

El sistema IR-n es un modelo de RP. En este apartado se presenta el modelo conceptual del sistema IR-n. Para ello se detallan las características del sistema IR-n en base a los siguientes conceptos:

1. Concepto de pasaje.
2. Cálculo de similitud entre pregunta y pasaje.
3. Cálculo de similitud entre documento y pregunta, en función de la similitud de los pasajes con respecto a ésta.
4. Aspectos adicionales que contempla el sistema IR-n.

4.2.1 El concepto de pasaje

Los modelos iniciales de RP unían la idea de pasaje a la de párrafo. Según el diccionario de la Real Academia de la Lengua Española, el párrafo es *"cada una de las divisiones de un escrito señaladas por letra mayúscula al principio del renglón y punto y aparte al final del trozo de escritura"*.

Además, otro elemento que los puede diferenciar dentro del texto es el uso de la sangría al principio del mismo, es decir la primera línea aparece desplazada a la derecha con respecto de las demás.

La definición de párrafo se puede ampliar a "un conjunto de frases –u oraciones– relacionadas que desarrollan un único tema.

Es una unidad intermedia, superior a la oración e inferior al texto, con valor gráfico y significativo" (Cassany, 1990).

Esta afirmación establece un vínculo entre párrafo y tema, que es el objetivo fundamental de localización de los sistemas de RP. Si se divide un documento en los temas que lo forman, se pueden realizar búsquedas por cada uno de ellos. Como consecuencia de esto, los primeros experimentos con pasajes se realizaron en base a los párrafos que forman los documentos.

No obstante, el utilizar los párrafos directamente como unidad de definición de los pasajes provoca que la colección de los mismos que se genera sea muy heterogénea en cuanto a tamaño. Además, no garantiza realmente que cada párrafo trate un tema diferente.

Por ello, las siguientes propuestas que contemplaron los modelos de RP utilizaban unidades de definición de pasajes basadas en los párrafos, aunque no utilizaran la equivalencia de un pasaje un párrafo:

- Los modelos semánticos forman los pasajes uniendo párrafos consecutivos en función del contenido de los mismos. Con ello, intentan solucionar el problema de que varios párrafos consecutivos traten el mismo tema. No obstante, como se ha demostrado, a pesar de ser un idea interesante, no se han conseguido algoritmos de segmentación que funcionen de forma similar a lo que se podría conseguir de forma manual (Hearst, 1994).
- Los modelos del discurso utilizan directamente los párrafos o secciones de los documentos, que están formados por varios párrafos. Estas propuestas únicamente funcionan razonablemente cuando se trata de documentos muy estructurados y hay cierta homogeneidad en el estilo de escritura utilizado por los autores de los documentos (Callan, 1994).
- Los modelos de ventana que utilizan elementos del discurso, suelen basarse en el párrafo, y definen los pasajes en base a un número de párrafos consecutivos de forma que el tamaño de los mismos se encuentre formado alrededor de un número determinado de bytes o palabras. De esta forma se consiguen colecciones de pasajes de tamaño muy homogéneo.

La idea de separar párrafo y pasaje aparece en los modelos de ventana que no utilizan la estructura del documento. Estos modelos cambian el párrafo por la palabra como unidad de definición de los pasajes. Esto permite una mayor flexibilidad a la hora de definir los pasajes y de adecuar el tamaño de los mismos a la colección de documentos y al tamaño de la pregunta del usuario. El hecho de que el sistema de RP sea flexible para poder adaptarse a estos aspectos tiene especial importancia en el rendimiento del sistema (Kaszkiel y Zobel, 2001). No obstante, los modelos que utilizan este esquema tienen como inconveniente la posible pérdida de la estructura en el pasaje. Por ejemplo, un pasaje de 100 palabras, puede contener una o dos frases incompletas dentro del mismo.

Entre la palabra y el párrafo existe una unidad con estructura que es la de la frase. La Real Academia de la Lengua Española define la frase como *"cualquier grupo de palabras conexo y dotado de sentido"*. En un texto, las frases vienen definidas por un símbolo de puntuación que define su fin.

El aspecto fundamental de la frase dentro de la problemática de la RI es la idea de que una frase tiene sentido, es decir una frase puede ser entendida por un usuario, lo cual le confiere interés como unidad de transmisión de información. Evidentemente, una única frase no tiene una identidad suficiente para determinar si un documento que la contenga es relevante respecto a determinado tema, aunque establece unos límites de forma que se pueda valorar el hecho de que términos de una consulta aparecen en la misma frase.

Dado que la frase en sí misma no tiene una entidad lo suficientemente completa para definir un pasaje, sí que es posible definir los pasajes como una serie de frases consecutivas dentro del texto.

El sistema IR-n utiliza la frase como unidad de definición de los pasajes. Así, en este sistema un pasaje se define como "un conjunto de frases consecutivas en un documento". El tamaño de cada pasaje, medido en número de frases que lo forman, es un parámetro que el sistema IR-n puede adecuar al ámbito de uso del mismo, con el objeto de optimizar el rendimiento del sistema.

Esta idea de frase como unidad de definición de los pasajes aporta una serie de ventajas a las de párrafo y palabra.

La idea de párrafo como elemento central tiene dos problemas fundamentales:

- Puede no disponerse de información acerca de la composición de los párrafos en el documento original.
- Los párrafos pueden utilizarse en ocasiones, más por motivos visuales que por la propia estructuración del documento.

La idea de número de palabras como elemento central presenta dos problemas:

- El número de palabras que se utilizan para describir un hecho determinado depende en gran medida del estilo de escritura que se utilice. Un mismo suceso es descrito con menos palabras en un documento de agencia de noticias que en un documento de una noticia de periódico. Si además el mismo hecho es tratado de forma literaria, el número de palabras utilizadas será sensiblemente mayor.
- Si se utilizan únicamente las palabras como elemento a considerar en la definición del pasaje, puede ocurrir que los pasajes considerados relevantes carezcan de estructura, al poder empezar y finalizar en cualquier parte del documento. Esto puede dificultar en gran medida la comprensión del texto recuperado.

Por otro lado, la frase como unidad presenta las siguientes ventajas sobre el uso de los párrafos o palabras:

- Una frase suele expresar una idea dentro del documento.
- En un texto con signos de puntuación se puede intuir casi con un cien por cien de efectividad, los límites que definen cada frase que forma un documento analizando exclusivamente su estructura superficial (Muñoz y Palomar, 1999).
- Las frases son unidades completas que permiten, bien mostrar a un usuario una información entendible, o bien que la salida del sistema de RI pueda ser la entrada de un sistema de tratamiento posterior (por ejemplo un sistema de BR). Por ello, el uso de frases mejora las propuestas que definen los pasajes que pueden empezar en cualquier palabra del documento.

- El uso de frases como unidad, permite poder comparar documentos de diferentes autores o colecciones que utilicen diferentes estilos literarios, como puedan ser noticias de agencia, noticias de periódico, enciclopedias, libros, etc.

Además, el uso de las frases como unidad de definición de los pasajes permite la misma flexibilidad en el momento de definir el tamaño de los mismos que los modelos de ventana que no utilizan la estructura del documento. El número de frases que forman un pasaje es un elemento parametrizable que puede depender de la colección de documentos utilizada, del tipo y tamaño de la pregunta así como de la finalidad o destino de la información recuperada.

Como resumen, la tabla 4.1 muestra las principales características de las diferentes unidades empleadas en la definición de pasajes. En la misma se puede comprobar que la frase utilizada como unidad de definición de los pasajes permite que éstos tengan estructura sintáctica y además facilita y flexibiliza la definición de los pasajes.

Unidad	Estructura sintáctica	Facilidad definición de los pasajes	Flexibilidad definición de los pasajes
Palabra	No	Alta	Alta
Frase	Si	Alta	Alta
Párrafo	Si	Baja	Baja

Tabla 4.1. Características de las unidades utilizadas en los modelos de RP

Formalmente la definición de los pasajes en el sistema IR-n se realiza de la siguiente forma:

- Dado un documento D formado por N frases f_i.

$$D = (f_i, ..., f_N) \tag{4.5}$$

- Considerando n como el tamaño de los pasajes medido en número de frases que lo forman, se definirían los siguientes pasajes P_i:

$$P_i = (f_{n*(i-1)}, ..., f_{min(n*i,N)}), i \in [1..(int(N/n) + 1)] \tag{4.6}$$

De esta definición cabe destacar:

- $n * (i - 1)$ fija la frase en la que empieza el pasaje P_i.
- El número de pasajes que se definen es $int(N/n) + 1$.
- Todos los pasajes están formados por n frases, exceptuando el último que puede estar formado por un número menor. Este último pasaje finaliza en la última frase del documento, y únicamente contendrá el mismo número de frases que el resto en el caso de que N sea múltiplo de n.

De la ecuación 4.6 se extraen los requisitos para poder realizar la definición de los pasajes en el sistema IR-n:

1. Es necesario conocer los límites de cada una de las frases que forman el documento.
2. Es necesario determinar cuál es el tamaño del pasaje a considerar, en base al número de frases que lo forman.

Según esta definición, suponiendo un tamaño de pasajes de 15 frases y un documento de 35 frases, se definen tres pasajes de la siguiente forma:

1. $P_1 = f_1..f_{15}$
2. $P_2 = f_{16}..f_{30}$
3. $P_3 = f_{31}..f_{35}$

4.2.2 Cálculo de similitud entre el pasaje y la pregunta

La medida de similitud permite cuantificar la semejanza entre un texto (ya sea un documento completo o un pasaje del mismo) y la pregunta. Estas medidas se basan fundamentalmente en los términos que comparten el texto y la pregunta así como en la importancia discriminatoria de cada término.

En un sistema RID para realizar estos cálculos de similitud, se define un documento D como un conjunto de pares de valores (d_i, n_i), en los cuales d_i sería el término y n_i el número de veces que aparece dicho término en el documento. El valor N representa el tamaño del documento, en cuanto al número de términos diferentes que lo forman.

$$D = ((d_1, n_1), (d_2, n_2), ..., (d_N, n_N)) \tag{4.7}$$

Una pregunta Q se define como un conjunto de pares de valores (q_i, m_i), en los cuales q_i sería el término y m_i el número de veces que aparece dicho término en la pregunta. El valor K indica el número de términos diferentes que forman la pregunta.

$$Q = ((q_1, m_1), (q_2, m_2), ..., (q_K, m_K)) \tag{4.8}$$

La medida de similitud entre Q y D se calcula en función de:

- El número de términos que comparten la pregunta y el documento.
- El número de apariciones en ambos de dichos términos compartidos.
- El valor discriminatorio o peso x_i del término dentro de la colección de documentos. Este peso x_i de un termino t_i se define en función del número de documentos de la colección en los que aparece dicho término.

Así, la medida de similitud se define de la siguiente forma:

$$sim(Q, D) = \Upsilon_{\forall_{i \in Q \wedge D}}((t_i, n_i, m_i, x_i), N) \tag{4.9}$$

En esta medida, Υ define un método para cuantificar el valor de la similitud entre documento y pregunta, en función de los parámetros indicados entre paréntesis.

En los sistemas de RP el cálculo de similitud entre pasaje y pregunta es igual al definido en 4.9, pero sustituyendo las apariciones de documento por las de pasaje, al que denominamos P. De esta forma la similitud entre un pasaje P de tamaño N' y una pregunta Q se define como:

$$sim(Q, P) = \Upsilon'_{\forall_{i \in Q \wedge P}}((t_i, n_i, m_i, x_i), N') \tag{4.10}$$

Los sistemas de RP, adicionalmente al cálculo de similitud de cada pasaje con respecto a la pregunta, deben obtener la similitud del documento con respecto a ésta. Este cálculo se obtiene en función del cálculo de la similitud obtenida por todos sus pasajes de la forma:

$$sim(Q, D) = \Phi_{\forall_{i:P_i \in D}}(sim(Q, P_i)) \tag{4.11}$$

Siendo Φ una función que cuantifica la similitud de una pregunta con respecto a un documento, en base a los valores de similitud de cada pasaje con respecto a la misma pregunta.

La mayoría de los sistemas de RP se basan en las mismas medidas definidas en los sistemas RID. Además, no existe en muchos de ellos una asignación directa del modelo que define la forma de segmentación del documento en pasajes y la medida de similitud utilizada. De hecho, Kaszkiel y Zobel (2001) combinan diferentes aproximaciones de segmentación de documentos con diferentes medidas de similitud. En la tabla 4.2 se muestran las medidas de similitud que son referenciadas en las definiciones de los modelos más conocidos de RP.

Modelo	Sistema	Medida de similitud
Discurso	Párrafos	Coseno
	Secciones	Coseno
Semántico	TextTiling	Coseno
Ventana	Pages	Coseno Pivotado
	Bounded Paragraphs	Coseno Pivotado
	Sliding Windows	Similar Coseno Pivotado sin Normalizar
	Arbitrary Passages	Coseno Pivotado

Tabla 4.2. Medidas de similitud utilizadas por los modelos de RP más representativos

Como se puede observar en esta tabla la mayoría de los modelos de RP usan medidas de similitud ya utilizadas por los modelos RID. Estas medidas contemplan la aplicación de normalización sobre la medida en función del tamaño del documento (o pasaje en este caso). La única excepción es el modelo *Sliding Windows* que define una nueva medida, en la que se destaca el hecho que no se aplica la normalización, ya que todos los pasajes están formados exactamente por el mismo número de palabras.

El planteamiento que se ha realizado en el sistema IR-n es diferente. En primer lugar cabe recordar los requisitos del sistema IR-n que son:

- Los pasajes se definen en base a un número de frases consecutivas.
- Este número de frases es un valor parametrizable que depende de factores externos al sistema como son la colección de documentos utilizada o la pregunta del usuario y/o el destino de la información recuperada.

Así la definición del documento varía sensiblemente con respecto a las realizadas anteriormente, ya que los pasajes están formados por frases consecutivas, y el tamaño del pasaje no se conoce de forma previa. Por ello, un documento se divide en una serie de frases de la forma:

$$D = (f_1, f_2, ..., f_m) \tag{4.12}$$

Para realizar los cálculos de similitud en el sistema IR-n es necesario conocer las apariciones de cada término de la pregunta en cada una de las frases del documento. Por ello la información necesaria de un documento se estructura de la siguiente forma:

$$D = \begin{pmatrix} d_1, (n_{11}, f_1) & (n_{12}, f_2) & .. & (n_{1l}, f_{l1}) \\ d_2, (n_{21}, f_1) & (n_{22}, f_2) & .. & (n_{2l}, f_{l2}) \\ .. & .. & .. & .. \\ d_N, (n_{N1}, f_1) & (n_{N2}, f_2) & .. & (n_{Nl}, f_{lN}) \end{pmatrix} \tag{4.13}$$

donde para cada uno de los términos d_i se conoce el par de valores (n_{ij}, f_j) que representa el número de veces que aparece dicho término en la frase f_j.

La definición de la pregunta Q no cambia con respecto a la realizada en los sistemas RID, siendo la siguiente:

$$Q = ((q_1, m_1), (q_2, m_2), ..(q_K, m_K)) \tag{4.14}$$

donde q_i es el término, m_i el número de veces que aparece dicho término en la pregunta y K el número de términos diferentes que forman la pregunta.

La definición de un pasaje $P_{a,b}$ se hace en función de un tamaño de pasaje Z, de tal forma que a indica la frase inicial y b la frase final del pasaje, y cumpliéndose que:

$$Z = b - a + 1 \tag{4.15}$$

La similitud de un pasaje $P_{a,b}$ y una pregunta Q dependerá de las apariciones de los términos de la pregunta en las frases de la a a la b de la forma:

$$sim(Q, P_{a,b}) = \Upsilon'_{\forall_{i \in Q \wedge fc/c \in [a,b]}}(t_i, n_i, m_i, x_i) \tag{4.16}$$

Hay que destacar dos aspectos fundamentales en esta definición:

1. Desaparece del cálculo el valor N que define el tamaño del documento o pasaje. Esto se debe a que no se aplican medidas de normalización con respecto al tamaño, ya que se considera que todos los pasajes tienen el mismo tamaño, indicado en base al número de frases que lo forman.
2. Para valorar el peso de cada término se utiliza la frecuencia del término en la colección de documentos en vez de la frecuencia del término en la colección de pasajes. Realizarlo de otra forma complicaría de forma notable el proceso de indexación y recuperación, ya que realmente no se conoce el valor del tamaño de los pasajes hasta el momento en el que se indique la pregunta. Además, existen estudios (Callan, 1994) que indican que no se aprecian mejoras al utilizar la frecuencia sobre pasajes.

El sistema IR-n define su medida de similitud a partir de la del coseno, pero aplicando los aspectos anteriormente indicados. De esta forma el sistema IR-n calcula la similitud de un pasaje P con respecto a una pregunta q de la siguiente forma:

$$sim(Q, P) = \sum_{t \in Q \wedge P} (w_{Q,t} \cdot w_{P,t} \cdot x_t) \tag{4.17}$$

donde:

- $w_{P,t}$ del número de apariciones del término t en el pasaje P.
- $w_{Q,t}$ del número de apariciones del término t en la pregunta Q.
- x_t del peso del término t.

La forma de calcular dichos valores es la siguiente:

$$w_{P,t} = \log_e(freq_{P,t} + 1) \tag{4.18}$$

$$w_{Q,t} = \log_e(freq_{q,t} + 1) \tag{4.19}$$

$$x_t = \log_e(\frac{N}{freq_t} + 1) \tag{4.20}$$

donde $freq_{Y,t}$ es el número de apariciones o frecuencia del término t en el pasaje o la pregunta Y, N es el número de documentos de la colección y $freq_t$ es el número de documentos diferentes que contienen el término t.

Esta forma de calcular la similitud de un pasaje con respecto a una pregunta se denomina $IR - n_{base}$ para diferenciarla de posteriores refinamientos que se han realizado a la misma y que se describen en la subsección 4.4.1 de este capítulo. La medida $IR - n_{base}$ fue la utilizada en los primeros experimentos realizados con el sistema IR-n, tanto en la conferencia CLEF-2001 (Llopis y Vicedo, 2001) como en una comparativa con otros modelos de RI (Llopis et al., 2002f).

4.2.3 Cálculo de la similitud del documento en función de la similitud de sus pasajes

Como ya se ha indicado en (4.11) todos los sistemas de RP calculan la similitud del documento en función de la similitud de sus pasajes, en los que la función Φ puede ser fundamentalmente, la suma de similitudes o la del pasaje de mayor similitud. Los experimentos de Hearst y Plaunt (1993) han sido contrastados en el sistema IR-n, obteniéndose siempre mejores resultados al utilizar la similitud del mejor pasaje como valor para la similitud del documento.

Este hecho indica, en general, que si un pasaje es relevante, el documento también lo es. Si un sistema de RP valora en lugar del concepto de mejor pasaje, el de suma de similitudes de los pasajes se comporta como un sistema RID, pero añadiendo conceptos de cercanía.

Otro aspecto importante es que la utilización de la función de mejor pasaje de cada documento, permite obtener el pasaje más relevante, facilitando así, tareas de proceso posteriores a la búsqueda.

Por todo ello, el sistema IR-n calcula la similitud del documento en base a la mayor similitud obtenida por sus pasajes:

$$sim(Q, D) = \max_{\forall i: P_i \in D} sim(Q, P_i) \tag{4.21}$$

4.2.4 Aspectos adicionales en la definición de pasajes en el sistema IR-n

Una vez determinada la frase como unidad empleada por el sistema IR-n en la generación de los pasajes y definida la medida utilizada para el cálculo de la similitud de los pasajes y documentos con respecto a una pregunta, quedan dos aspectos por analizar que son los siguientes:

- Grado de solapamiento de los pasajes
- Momento en el que se definen los pasajes

Uso de pasajes solapados. Como ya se estudió en el capítulo anterior, todas las definiciones de pasajes pueden no ser perfectas. Por ello puede ocurrir que los términos de la pregunta aparezcan dispersos en más de un pasaje, con lo que se podrían descartar documentos relevantes al realizar un estudio basado en pasajes. Este problema se puede evitar con el uso de pasajes solapados, ya que éstos permiten que más de un pasaje comparta el mismo fragmento de texto del documento.

Aprovechando que el sistema IR-n emplea la frase como unidad de definición de los pasajes, se toma también ésta para la definición del solapamiento de pasajes. El grado de solapamiento (G_{sol}) en el sistema IR-n indica el número de frase de un pasaje a partir del cual comienza la definición del siguiente pasaje. Las principales características de este valor son las siguientes:

1. G_{sol} debe ser siempre menor que el tamaño del pasaje. Si tuviera el mismo valor indicaría que no se utiliza solapamiento.
2. Cuanto menor es el valor de G_{sol} mayor es la cantidad de texto que comparten dos pasajes consecutivos.
3. Como consecuencia de la anterior, cuanto menor es el valor de G_{sol}, se definirán un mayor número de pasajes en un documento.

La utilización de solapamiento implica redefinir el concepto de pasaje en el sistema IR-n, quedando de la siguiente forma:

- Dado un documento D formado por N frases.

$$D = f_1..f_N \tag{4.22}$$

- Considerando n como el tamaño de los pasajes medido en número de frases que lo forman.
- Dado un grado de solapamiento G_{sol}
- Se definirían los siguientes pasajes a partir del documento D

$$P_i = f_{G_{sol}*(i-1)+1}, \cdots, f_{\min(G_{sol}*(i-1)+n,N)}, i \in [1..N/G_{sol}-1] \tag{4.23}$$

Dada esta definición y suponiendo un tamaño de pasajes de 15 frases, un grado de solapamiento de grado 10 y un documento de 35 frases, la generación de pasajes se realizaría de la siguiente forma:

1. $P_1 = f_1..f_{15}$
2. $P_2 = f_{11}..f_{25}$
3. $P_3 = f_{21}..f_{35}$

El solapamiento de los pasajes tiene como ventaja su incremento de eficacia en la recuperación de documentos. Sin embargo, se incrementa el tiempo de respuesta del sistema, en mayor medida cuando el grado de solapamiento es menor, ya que se incrementa el número de pasajes a evaluar.

No obstante, como se demostrará en la fase de experimentación realizada (capítulo 5), utilizar un grado de solapamiento menor mejora sensiblemente los resultados del sistema y no incide excesivamente en el tiempo total de búsqueda.

El hecho que el solapamiento no incremente de forma notable el coste total de la búsqueda se debe principalmente a dos motivos:

1. El sistema IR-n no evalúa todos los pasajes que forman el documento, ya que se pueden evitar el cálculo de similitud de algunos de ellos. En primer lugar el primer pasaje que se evalúa es aquel que se inicia en la primera frase del documento en la que aparece un término de la pregunta. Esto se debe a que ningún pasaje que comience en una frase anterior puede

obtener un valor de similitud superior a este primer pasaje, dada la forma en la que se ha definido la medida de similitud en el sistema IR-n.

Dados dos pasajes $P_{a,b}$ y $P_{c,d}$, siendo a, b, c y d números de frase que definen los límites de los pasajes y cumpliéndose:

$$a \leq c \leq b \leq d \tag{4.24}$$

o sea, ambos pasajes comparten las frases que van desde la número c hasta la indicada por b. Si no existe ningún término de la pregunta que aparezca en las frases de la a a la c:

$$\forall t_i \in Q \rightarrow t_i \notin P_{a,c} \tag{4.25}$$

Por ello, la similitud del pasaje $P_{a,c}$ es igual a 0, por lo que, siempre la similitud del pasaje $P_{a,b}$ será menor o igual a la del pasaje $P_{c,d}$.

$$sim(Q, P_{a,b}) \leq sim(Q, P_{c,d}) \tag{4.26}$$

Por el mismo motivo, el último pasaje que se evalúa es aquél que finaliza en la última frase del documento en la que aparece un término de la pregunta.

Estas mismas conclusiones pueden extenderse a pasajes que no se encuentren en los extremos del documento, es decir, a pasajes interiores. Dado un grado de solapamiento G_{sol}, si un pasaje en sus primeras G_{sol} frases no contiene ningún término de la pregunta también se puede obviar su evaluación sin modificar por ello el resultado final. Por ejemplo, si G_{sol} es igual a 1, no es necesario evaluar aquellos pasajes cuya primera frase no contiene ningún término de la pregunta.

Debido a todo ello se reduce notablemente el número de pasajes a evaluar, y en consecuencia el utilizar un grado de solapamiento pequeño no incide de la misma forma que si se evalúan todos los pasajes posibles en un documento.

2. Otro aspecto es la forma en la que se ha implementado el sistema, que se basa en cargar en memoria principal toda la información sobre las apariciones de palabras en los documentos. Este aspecto se describe con mayor profundidad en la subsección 4.3.3 dentro de este capítulo. De esta forma, el

proceso de segmentación se realiza en tiempo de ejecución y sobre estructuras de datos albergadas en memoria principal. Considerando que los factores que más influyen en el tiempo de proceso están relacionados con los tiempos de acceso a disco, este ligero incremento al procesar un mayor número de pasajes sobre memoria principal, no incrementa de forma ostensible el tiempo final necesario para procesar todos los pasajes.

Por ello, el sistema IR-n utiliza un grado de solapamiento (G_{sol}=1) que es con el que mejor rendimiento se obtiene, como se mostrará en la experimentación realizada (capítulo 5).

Momento en el que se definen los pasajes. Como ya se ha indicado y se demostrará en la experimentación realizada, el tamaño óptimo de los pasajes depende de la colección de documentos que se utiliza y del tipo de la pregunta. La colección de documentos que se utiliza se conoce en el momento de la indexación. No ocurre de la misma forma con la pregunta ya que ésta se conoce en el momento de la búsqueda.

El hecho que el tipo de la pregunta influye en el tamaño del pasaje con el que se obtienen mejores resultados, se demuestra en los experimentos realizados en el capítulo 5 y en Kaszkiel y Zobel (2001).

Otro aspecto importante es la complejidad temporal y espacial de los procesos de indexación y búsqueda según el momento en el que se definen los pasajes. El realizar la división de los pasajes en el momento de la indexación tiene como principal ventaja la disminución del tiempo del proceso de búsqueda. Como principal inconveniente, cabe citar, el hecho de que el tamaño de los pasajes no se podrá definir en función de la pregunta que realice el usuario. Otro inconveniente que se produce en el caso de que se utilicen pasajes solapados, como en el sistema IR-n, es que se incrementa la cantidad de información a almacenar, ya que se debe contemplar, para cada palabra, todos los pasajes en los que aparece.

El realizar la división de los documentos en pasajes en el momento de la búsqueda otorga una flexibilidad importante al siste-

ma, ya que puede decidir el tamaño de los pasajes en el momento de procesar la pregunta. Otra ventaja es que permite reducir la cantidad de información que se almacena. Como inconveniente cabe citar el hecho de que añade al proceso de búsqueda de documentos relevantes, la tarea de la descomposición del documento en pasajes.

Para dotar al sistema IR-n de la mayor flexibilidad posible se ha adoptado la decisión de que la división de los pasajes se realice en tiempo de búsqueda. En la sección 4.3, dedicada a la arquitectura del sistema, se indicará la forma en la que se ha implementado, de forma que permite disminuir en lo posible la complejidad del proceso de búsqueda.

Como resumen del enfoque utilizado por los principales modelos de RP para la definición de pasajes, se describe en la tabla 4.3 las principales características de los mismos. Como se puede observar en esta tabla, el modelo IR-n se halla encuadrado en los modelos de ventana y se diferencia de las otras propuestas, fundamentalmente, en la unidad que utiliza para definir los pasajes. En la misma, se puede comprobar que el sistema IR-n es el único que utiliza las frases como unidad para definir los pasajes, mientras el resto de propuestas se basan bien en el uso de párrafos, palabras o una combinación de los mismos. Por otro lado, el sistema IR-n comparte con los modelos *Arbitrary Passages* y *Sliding Windows* tanto el hecho de utilizar solapamiento como el de realizar la definición de los pasajes en tiempo de búsqueda.

4.3 Arquitectura del sistema IR-n

Una vez estudiado el modelo conceptual del sistema IR-n, cabe plantearse su implementación. Los objetivos fundamentales que se fijan con respecto al sistema son los siguientes:

1. **Eficacia.** La eficacia del sistema se va a valorar a nivel de las medidas más conocidas y que posiblemente reflejen mejor los resultados que obtiene un sistema de RI. Éstas son las de cobertura y precisión.

Modelo	Sistema	Unidad	Solapamiento	División
Discurso	Párrafos	Párrafos	No	Indexación
	Secciones	Marcas HTML	No	Indexación
Semántico	TextTiling	Párrafo	No	Indexación
Ventana	Pages	Párrafo + Caracteres	No	Indexación
	Bounded Paragraphs	Párrafo + Palabras	No	Indexación
	Sliding Windows	Palabras	Sí. Cada n/2 palabras. Siendo n el tamaño de los pasajes	Búsqueda
	Arbitrary Passages	Palabras	Sí. Cada 25 palabras	Búsqueda
	IR-n	Frases	Sí. Cada frase	Búsqueda

Tabla 4.3. Definición de pasajes en modelos de RP más representativos

2. **Eficiencia.** La idea de eficiencia se plasma en que el sistema propuesto debe ser aplicado en tiempo real.
3. **Flexibilidad.** El sistema debe permitir tratar cualquier tipo de documento, tanto aquéllos que se encuentran en texto plano como aquellos que utilizan algún formato (SGML, HTML, etc.).
4. **Conectividad.** El sistema debe permitir que su salida pueda ser utilizada por sistemas de tratamiento posterior de la información recuperada, como por ejemplo, sistemas de BR.

Para la definición de la arquitectura del sistema IR-n se va a partir del esquema general definido en la figura 2.5, del capítulo 2, donde se comentaban los principales módulos que componen un sistema de RI: preproceso de textos, indexación, gestión de la pregunta y búsqueda. Además, para dotar de mayor flexibilidad al sistema se ha incorporado un nuevo módulo denominado *conversión de documentos*, el cual tiene como objetivo convertir los documentos originales a un formato estándar. De esta forma se evita que los módulos de preproceso y/o indexación deban ocuparse de estas tareas. La arquitectura general del sistema IR-n se muestra en la figura 4.1.

Así, los módulos que componen el sistema IR-n son los siguientes:

1. Módulo de conversión de documentos.

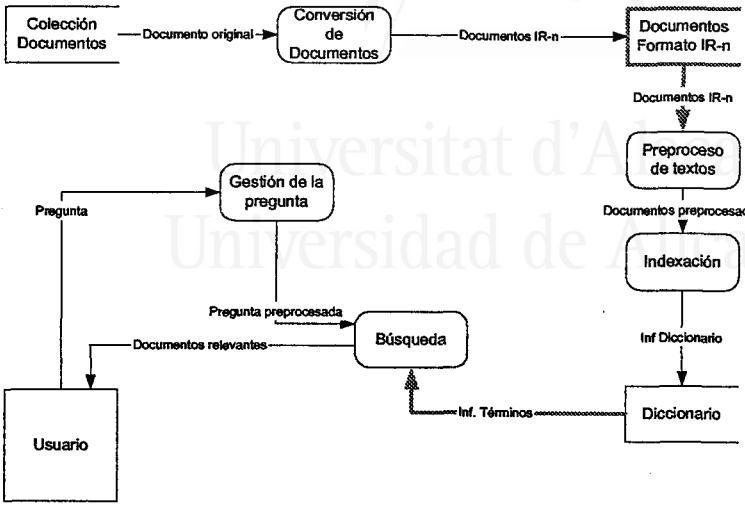


Figura 4.1. Arquitectura general del sistema IR-n

2. Módulo de preproceso de textos.
3. Módulo de indexación.
4. Módulo de gestión de la pregunta.
5. Módulo de búsqueda.

Todos estos módulos se describen de forma más detallada en las siguientes subsecciones.

4.3.1 Módulo de conversión de documentos

Los documentos que pueden formar parte de las colecciones a indexar pueden hallarse en diferentes formatos. El tratamiento de estos formatos por parte del proceso de indexación incrementaría la complejidad del mismo. Por ello, se define un formato único, denominado formato IR-n, que será el contemplado por el módulo de indexación. Por último, se definen una serie de mecanismos que permiten convertir los documentos a este formato (ver figura 4.2).

La conversión de documentos facilita el trabajo con las colecciones TREC y CLEF que se hallan en formato SGML.

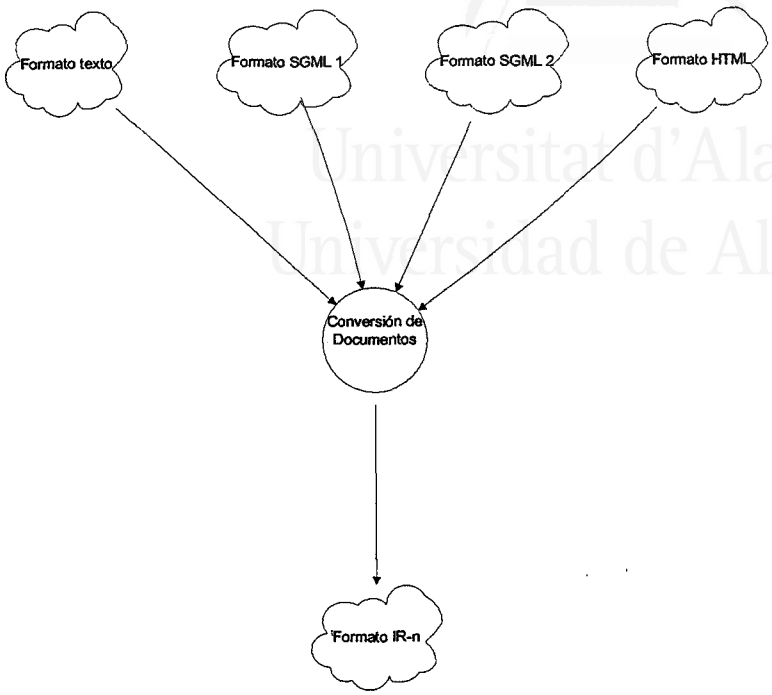


Figura 4.2. Conversión de documentos al formato IR-n

Como se ha indicado, el sistema IR-n define el documento en función de las frases que lo forman. En consecuencia, se requiere conocer además del identificador del documento, los límites de todas las frases que contiene. El formato IR-n emplea dos etiquetas que permiten identificar sin ambigüedad estos dos aspectos: *DOCNO* y *ENDF*.

<*DOCNO*> Es una etiqueta que antecede al nombre del documento. Este nombre debe ser único, ya que es el que permitirá identificar al documento y diferenciarlo del resto.

<*ENDF*> Es una etiqueta que delimita el final de una frase.

En (24) se muestra un ejemplo de texto convertido a IR-n. En éste, la etiqueta *DOCNO* marca el inicio del documento y precede al nombre del mismo. El resto del documento está formado por una serie de palabras, y los finales de frase se indican mediante la etiqueta *ENDF*.

(24) <DOCNO>EFE19940101-00001
GUINEA-OBIANG PRESIDENTE SUGIERE RECHAZARA AYUDA EXTERIOR CONDICIONADA <ENDF>
Malabo, 31 dic (EFE).- <ENDF>
El presidente de Guinea Ecuatorial, Teodoro Obiang Nguema, sugirió hoy, viernes, que su Gobierno podría rechazar la ayuda internacional que recibe si ésta se condiciona a que en el país haya "convulsiones políticas". <ENDF>
En su discurso de fin de año, Obiang afirmó que "la democracia pluralista no es sinónimo de desorden ni de convulsiones", y el pueblo guineano "no tiene necesidad de sufrir convulsiones políticas ni desórdenes sociales", y "si de estas convulsiones depende la ayuda que nos han prometido, preferimos rechazar dicha ayuda para seguir valientemente nuestro proceso de desarrollo sin ayuda extranjera". <ENDF>

La conversión de documentos a formato IR-n se realiza en dos fases:

1. **Lectura y filtrado del documento original.** Tiene como objetivo detectar las palabras que forman parte del documento según el formato original del mismo. Dada la importancia de este aspecto por el uso del sistema IR-n en las conferencias del TREC y CLEF, se definió un módulo específico de conversión de documentos SGML al formato IR-n.
2. **Detección de las frases en el documento.** El módulo de descomposición de frases utilizado en el sistema IR-n, se basa en el definido en Muñoz y Palomar (1999), con un porcentaje de frases correctamente segmentadas cercano al 100%.
   Los aspectos fundamentales considerados para determinar el límite de las frases de un documento son los siguientes:
   - Se estudian las palabras alrededor de los caracteres que pueden definir el final de la frase (., ?, !, ;).
   - Si la palabra siguiente a estos caracteres no comienza por mayúscula no se considera como final de frase.

- En caso contrario, se estudia la palabra que aparece delante del carácter:
  - Si esta palabra aparece en un diccionario de abreviaturas no se considera final de frase.
  - Si el carácter aparece entre números no se considera final de frase.
  - En cualquier otro caso se considera el comienzo de una nueva frase con aquellas palabras situadas a la derecha del carácter.

4.3.2 Módulo de preproceso de textos en el sistema IR-n

El módulo de preproceso de textos tiene como objeto preparar la información contenida en los documentos originales para que pueda servir de entrada al módulo de indexación.

Recibe como entrada la salida del módulo de conversión de documentos anterior, con el objetivo de tratar cada uno de los términos de la colección, para determinar si es un término a indexar y en caso afirmativo generar el *stem* del mismo. Los pasos que realiza este proceso son los mismos que fueron presentados en la sección 2.2.1:

1. Filtrado de caracteres no alfabéticos o numéricos
2. Conversión de palabras a minúsculas
3. Filtrado de palabras de parada
4. Filtrado de palabras adicionales no relevantes
5. Generación de *stems*

4.3.3 Módulo de indexación

El módulo de indexación construye los diccionarios e índices a partir de los ficheros en formato IR-n generados por el módulo anterior. Como se ha indicado en la subsección 4.2.2 para cada término indexable (es decir, aquél que se encuentre en al menos un documento y no sea una *palabra de parada*), se debe conocer aquellos documentos (y las frases dentro de cada uno de ellos) que lo contienen. Así, para cada término indexable (en adelante término) se necesita conocer la siguiente información:

- El número de documentos en los que aparece[1].
- El numero de veces que aparece en cada documento.
- Para cada aparición:
  - Número de frase en la que se encuentra.

Estructuras del proceso de indexación. La información que se almacena para cada término es de tamaño variable, ya que cada uno de ellos aparece en un número diferente de frases. Además, es importante garantizar un acceso rápido para localizar la información referente a cada uno de los términos de la pregunta, que son el punto de entrada del módulo de búsqueda. Así, esta información se almacena en dos estructuras de datos fundamentales que son:

- **Los ficheros de información.** Los ficheros de información contendrán toda la información referente a las apariciones de cada palabra en los diversos documentos y frases. Toda esa información se grabará en un fichero de forma secuencial, pero en ficheros que permitan el acceso directo a cualquier parte del mismo.
- **Diccionario.** El diccionario es una estructura indizada que permite el acceso rápido mediante el término y lo relaciona con la dirección física del fichero de información que contiene los datos referentes a dicho término.

Las relaciones entre ambas estructuras se muestran en la figura 4.3.

La información que gestionan estas dos estructuras es la siguiente:

Diccionario. Cada entrada del diccionario contiene la siguiente información para cada uno de los diferentes términos que se hallan en al menos un documento de la colección:

- *Stem.*
- Dirección física del fichero de información que contiene las apariciones del término.
- Número de documentos en los que aparece el término.

[1] Este valor permite calcular el peso del término en la colección

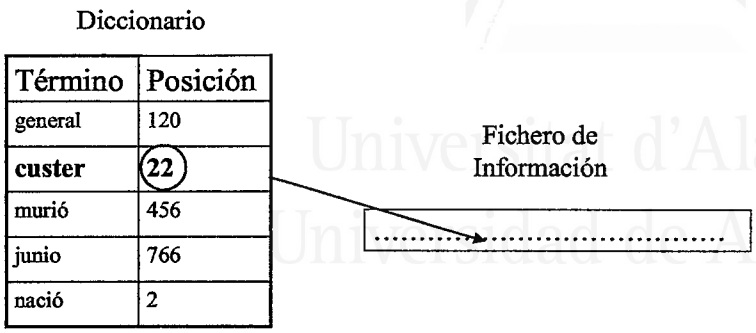


Figura 4.3. Diccionario y ficheros de información

Dado que la entrada a esta estructura es por término, éste será la clave de acceso a la misma. Para que el tiempo de acceso a la información de cada término sea lo más pequeño posible, se han utilizado tablas de dispersión (Hash). Estas tablas se utilizan en programación para el almacenamiento y recuperación rápida de claves únicas y valores asociados (Deitel y Deitel, 1999). Esta estructura permite únicamente la asociación *uno a uno* entre clave y valor. Éste es el requerimiento de la estructura Diccionario que únicamente tiene una entrada por término.

La principal ventaja de esta estructura es su rapidez, ya que está albergada en la memoria principal del ordenador. Como inconveniente cabe citar que se requiere realizar inicialmente un proceso de carga de la tabla (albergada en unidades de almacenamiento secundarias) a la memoria principal.

Ficheros de información. El acceso al diccionario a través de un término permite conocer la dirección física del fichero de información, que contiene toda la información referente al mismo. El soporte que se utiliza para definirlos son los ficheros que utiliza el lenguaje C++. Éstos permiten el uso de los mismos, tanto de forma secuencial como directa a través de la dirección física.

Si la colección de documentos es pequeña, se utiliza un único fichero de este tipo para toda la colección. Pero en el caso de que la colección de documentos tenga un tamaño considerable, se produciría un fichero de gran tamaño, poco manejable. Además, se complicaría notablemente el proceso de construcción del mismo (Kaszkiel et al., 1999), ya que para crear dicho fichero se deben mantener una serie de estructuras en memoria central, que debería ser de un tamaño considerable para almacenar tal cantidad de información.

El planteamiento original que utiliza el sistema IR-n, es disponer de un fichero de información para cada una de las letras del abecedario, así como uno adicional para los no alfabéticos. Cada uno de ellos contendrá todos los términos que empiecen por una letra determinada. Por ejemplo, el primero (letra a) contendrá información de todas las palabras que empiecen por la letra "a". Además, se crea un fichero para todos aquellos elementos que no empiecen por una letra y lo hagan por cualquier otro carácter no alfabético. Esta gestión del almacenamiento presenta una serie de ventajas importantes:

1. Disminuye el tamaño de los ficheros de información utilizados, facilitando su uso por parte del sistema operativo y los operadores del mismo.
2. Define de forma sencilla en qué fichero se halla la información referente a cada término, ya que se localiza por la inicial del mismo.
3. En sistemas multidisco o distribuidos, permite balancear de forma sencilla los accesos a los diferentes discos, disminuyendo los tiempos de acceso previos.
4. Disminuye el tamaño del registro que utiliza la tabla hash. Esto se consigue ya que las direcciones físicas máximas de los ficheros son de menor tamaño, que si se utilizase un único fichero.

En la figura 4.4 se muestra el funcionamiento de estas dos estructuras. El diccionario contiene una entrada para cada término. Para recuperar la información asociada a dicho término se utiliza la inicial del mismo para determinar el fichero a acceder, y

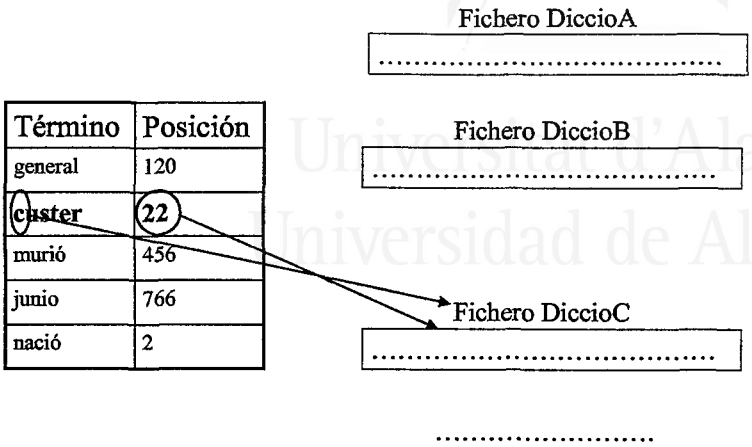


Figura 4.4. Diccionario y ficheros de información

la dirección donde se halla la información que se encuentra en el diccionario. Dentro de cada fichero se almacenará la siguiente información:

- Para cada documento en los que aparece el término:
  - Número del documento[2].
  - Número de veces que aparece el término en el documento.
  - Para cada aparición en cada documento.
    - Número de frase en la que aparece.

Dados los documentos descritos en (25):

(25) $< DOCNO >$ 1
El proceso de paz palestino-israelí se llevo a cabo en secreto $< ENDF >$
Sharon habló con Arafat $< ENDF >$

[2] Si el documento viene referenciado por un nombre, se utiliza una estructura adicional que relaciona este número con el nombre del documento

$< DOCNO > 2$
El líder israeli Sharon habló de paz con el líder palestino Arafat. $< ENDF >$

El preproceso de ambos documentos generarían las dos listas de términos a indexar, una para cada uno de los documentos. Éstas se pueden ver en (26):

(26) **Documento 1. Frase 1**
proces
paz
palestin
israeli
llev
cabo
secret
Documento1. Frase 2
sharon
habl
arafat
Documento 2. Frase 1
lider
israeli
sharon
habl
paz
lider
palestin
arafat

Como se puede comprobar, ya aparece el concepto de frase en el momento de localizar las apariciones de cada uno de los términos.

El diccionario que se genera para estos documentos se muestra en la tabla 4.4. En esta tabla se puede ver cómo el término "paz" aparece en dos documentos, mientras el término "secret" únicamente aparece en uno. Además, asociado a cada uno de estos dos términos se halla la dirección física del fichero que contiene la información completa de las apariciones de dichos términos.

Palabra	Dir física	Num apariciones
arafat	0	2
cabo	0	1
habl	0	2
israeli	0	2
llev	0	1
lider	4	1
palestin	0	2
proces	8	1
paz	12	2
secret	0	1
sharon	4	2

Tabla 4.4. Ejemplo de visualización del diccionario

Los ficheros de información almacenan la misma según se muestra en la tabla 4.5.

Por ejemplo, la palabra "paz" en el diccionario tenía asociada la dirección física 12. Con esta dirección y la inicial de la palabra (en este caso la letra "p"), se accedería al fichero "DiccioP" donde en la posición física 12, aparece toda la información antes especificada de la palabra "paz". Cabe añadir que realmente el término no se graba en el fichero, simplemente se muestra en la tabla para facilitar su comprensión.

A partir de estas estructuras se puede conocer de forma rápida en qué documentos se hallan los términos de la pregunta.

DiccioA				
Palabra	Num Apariciones	Num Doc	Num Apariciones	Frase
arafat	2	1	1	2
		2	1	1
DiccioC				
Palabra	Num Apariciones	Num Doc	Num Apariciones	Frase
cabo	1	1	1	1
DiccioH				
Palabra	Num Apariciones	Num Doc	Num Apariciones	Frase
habl	2	1	1	1
		2	1	1
DiccioI				
Palabra	Num Apariciones	Num Doc	Num Apariciones	Frase
israeli	2	1	1	1
		2	1	1
DiccioL				
Palabra	Num Apariciones	Num Doc	Num Apariciones	Frase
lider	1	2	2	1
				1
llev	1	1	1	1
DiccioP				
Palabra	Num Apariciones	Num Doc	Num Apariciones	Frase
palestin	2	1	1	1
		2	1	1
proces	1	1	1	1
paz	2	1	1	1
		2	1	1
DiccioS				
Palabra	Num Apariciones	Num Doc	Num Apariciones	Frase
secret	1	1	1	1
sharon	2	1	1	2
		2	1	1

Tabla 4.5. Ejemplo de visualización de ficheros índice

4.3.4 Módulo de gestión de la pregunta

Este módulo es muy similar al de preproceso de los documentos. Tiene como objetivo seleccionar los términos de la pregunta que se van a valorar en el proceso de recuperación. Los pasos que realiza son los siguientes:

1. Filtrado de caracteres no alfabéticos o numéricos
2. Conversión de palabras a minúsculas
3. Filtrado de palabras de parada
4. Filtrado de palabras adicionales no relevantes
5. Definición del término a indexar

El tratamiento realizado es idéntico al citado en la fase de preproceso de los documentos. Un ejemplo del tratamiento de la pregunta realizado sería el definido en (27).

(27) Entrada:
Reunión de Sharon con el líder palestino.
Salida:
reunion sharon lider palestin

4.3.5 Módulo de búsqueda

El módulo de búsqueda tiene como objetivo ordenar los documentos de la colección en función de su relevancia hacia una pregunta o tema determinado. Este módulo realiza tres tareas:

1. Selección de documentos
2. Cálculo de la relevancia de cada documento
3. Visualización de los resultados

Selección de documentos Este módulo tiene como objetivo determinar en qué documentos, y dentro de éstos, en qué frases aparecen cada uno de los términos de la pregunta. Los pasos que realiza para cada uno de los términos que forman la pregunta son los siguientes:

- Búsqueda en el diccionario de cada término.
  - Si no aparece se desecha el término.
  - Si aparece:
    - Recuperar la dirección donde se almacena su información.
    - Acceso al fichero de datos para recuperar todas las apariciones del término.

Palabra	Num documentos	Lugares
lider	1	(Doc 2, Frase 1) (Doc 2, Frase 1)
palestin	2	(Doc 1, Frase 1) (Doc 2, Frase 1)
sharon	2	(Doc 1, Frase 2) (Doc 2, Frase 1)

Tabla 4.6. Ejemplo de selección de documentos

Así el proceso de la pregunta descrita en (27), generaría la estructura que se muestra en la tabla 4.6. En esta tabla se muestra la información que se recupera para cada uno de los términos que forman la pregunta, indicando en qué documentos y en qué frases aparecen. Se puede observar que la palabra "reunión" se descarta y no se utiliza, ya que no existe ninguna entrada en el diccionario para la misma. Esto se debe a que ningún documento de la colección contiene dicha palabra, por lo que no va a ser utilizada en el proceso de búsqueda.

Cálculo de relevancia de los documentos Esta tarea tiene como objetivo calcular la relevancia o similitud con respecto de la pregunta de todos aquellos documentos que contienen al menos una aparición de un término de la pregunta. Los pasos que se realizan para cada uno de los documentos seleccionados en la tarea anterior son los siguientes:

1. *Generación de la matriz de apariciones.* Se genera una matriz en la que se indica para cada palabra y frase los lugares en los que aparece dentro de cada uno de los documentos. Un ejemplo de la misma para los documentos 1 y 2, descritos en (26), se muestra en las tablas 4.7 y 4.8 respectivamente. En la tabla 4.7 se puede ver que en la frase 1, el término "líder" aparece una vez y en la frase 2 el término "sharon" aparece una vez para el documento 1.

Frase	lider	palestin	sharon
1	1	0	0
2	0	0	1

Tabla 4.7. Ejemplo de matriz de apariciones para documento 1

Por otra parte, la tabla 4.8 muestra que en el documento 2, aparece en la frase 1, aparecen 2 veces el término "líder", y 1 vez los términos "palestin" y "sharon".
A partir de estos datos, el sistema debe calcular la relevancia de cada uno de estos documentos, en función de la relevancia de los pasajes que la forman.

Frase	lider	palestin	sharon
1	2	1	1

Tabla 4.8. Ejemplo de matriz de apariciones para documento 2

2. *Obtención de puntuación para cada pasaje.* Esta tarea calcula la relevancia de cada pasaje con respecto a la pregunta. La realización de este cálculo depende de tres aspectos fundamentales: la medida de similitud empleada, el tamaño de cada pasaje y el grado de solapamiento definido.
Supongamos que la tabla 4.9 muestra la matriz de ocurrencia de un documento con respecto a una pregunta formada por tres términos (pal1, pal2 y pal3).

Frase	pal1	pal2	pal3
1	0	0	0
2	2	1	1
3	2	1	1
4	2	1	1
5	0	0	1
6	0	0	0

Tabla 4.9. Ejemplo de matriz de apariciones para un documento

Suponiendo además, un tamaño de pasaje fijado en 2 frases y un grado de solapamiento de una frase, los pasos que se realizan para valorar los pasajes son los siguientes:

a) Definición de los pasajes a valorar, teniendo en cuenta que el primer pasaje empieza en la primera frase donde al menos aparece una de las palabras de la pregunta y el último pasaje acaba donde se halla la última aparición en el documento de una de las palabras de la pregunta. Se definen tres pasajes: pasaje 1 (frases 2 y 3), pasaje 2 (frases 3 y 4) y pasajes 3 (frases 4 y 5).
b) Cálculo de relevancia de cada pasaje. La similitud de cada pasaje con respecto de la pregunta se obtiene utilizando la medida definida en la subsección 4.2.2. Esta medida depende del peso de cada término de la pregunta y el número de apariciones del mismo en el pasaje.

3. *Asignación de la puntuación al documento.* Esta tarea tiene como objetivo el otorgar una puntuación de similitud al documento en función de las similitudes obtenidas por cada uno de los pasajes. El sistema IR-n otorga al documento el valor máximo de similitud alcanzado por los pasajes que lo conforman tal como se describe en la subsección 4.2.3.
4. *Ordenación de la lista de documentos.* El último paso de este módulo consiste en ordenar todos los documentos en función de la puntuación que ha obtenido cada uno de ellos.

Visualización de los resultados Este módulo tiene como objetivo mostrar una serie de datos a partir de la información generada en la fase anterior. En función de cuál sea el objetivo final de uso del sistema este módulo generará dichos datos según un determinado formato.

Si el objetivo es indicar la relevancia de cada documento se deberá mostrar el código que identifique al mismo junto a la relevancia calculada para el mismo. Si además se desea mostrar el contenido del pasaje más relevante de cada documento, bien para un sistema de búsqueda interactiva, bien para un sistema de BR, el sistema IR-n suministrará a este módulo el inicio del pasaje más relevante. Una vez conocido este dato, el módulo de visualización deberá acceder a los ficheros originales y seleccionar el pasaje más relevante en función del nombre de documento, la frase en la que empieza y el tamaño del pasaje.

Cuando se ha utilizado el módulo de visualización para las tareas de BR o simplemente para mostrar al usuario el pasaje más relevante, siempre se ha incorporado al pasaje la frase inmediatamente anterior (exceptuando el caso en el que el pasaje empiece en la primera frase). Esto se debe, a que los pasajes empiezan con frases en las que aparezca al menos un término de la pregunta. Se ha comprobado experimentalmente que la frase inmediatamente anterior tiene una considerable probabilidad de aportar información útil para la comprensión del pasaje, ya que es un antecedente de la primera frase del pasaje. Esta característica ha sido especialmente útil en la versión interactiva del sistema IR-n, presentada en el iCLEF-2002 (Llopis et al., 2002h), facilitando al usuario la

determinación de la relevancia de un documento estudiando sólo el pasaje más relevante junto con la frase inmediatamente anterior al mismo.

Una vez definido e implementado el sistema IR-n, se procedió a realizar una serie de experimentos para verificar su eficiencia y eficacia. Los experimentos más relevantes realizados se describen en el capítulo 5. A partir de los resultados obtenidos en dicha experimentación, se detectaron una serie de detalles del modelo IR-n que eran susceptibles de ser mejorados. Esto hizo que se definieran una serie de refinamientos al sistema originalmente planteado, que permitieron mejorar los resultados obtenidos por el sistema y le dotaron de mayor flexibilidad para adecuarse a diferentes tipos de preguntas. Estos refinamientos se detallan en la sección siguiente.

4.4 Refinamiento del sistema IR-n

En secciones anteriores se ha presentado el modelo conceptual del sistema IR-n, así como las características fundamentales de su arquitectura. El sistema así descrito fue utilizado en las ediciones del CLEF-2001 (Llopis y Vicedo, 2001) y TREC-2001 (Vicedo et al., 2002). El estudio exhaustivo de los resultados obtenidos permitió definir unas modificaciones al sistema originalmente planteado. Estos aspectos son los siguientes:

1. **Medidas de cercanía.** Incluir en la medida de similitud factores que permitan considerar la cercanía en la que aparecen los términos de la pregunta en el pasaje.
2. **Incorporación de técnicas de expansión de la pregunta.** Incluir en el sistema técnicas de expansión de la pregunta que permitan localizar documentos que siendo relevantes no contienen exactamente los mismos términos de la pregunta.
3. **Tratamiento en el sistema IR-n de las preguntas largas formadas por más de una frase.** Contemplar una descomposición de la pregunta de forma previa a su proceso.

Estos aspectos se tratan en los siguientes apartados.

4.4.1 Medidas de cercanía

En este sentido la modificación de la medida $IR{-}n_{base}$, definida en 4.17, se ha planteado fundamentalmente en el hecho de valorar qué términos, que se encuentren de forma consecutiva en la pregunta, aparecen en la misma frase del documento.

Un ejemplo de este caso es la pregunta utilizada en el CLEF "vacas locas en Europa", (que una vez eliminadas las *palabras de parada* se convertiría en "vacas locas europa". Es evidente que el hecho de que las palabras "vacas" y "locas" aparezcan de forma consecutiva en la misma frase de un documento, puede determinar que éste sea más relevante que en el caso que aparezcan en frases diferentes, aunque sea en el mismo pasaje.

A partir de esta hipótesis se ha definido una modificación a la medida $IR - n_{base}$. Esta medida se denomina $IR - n_{prox}$ y es un refinamiento de la anterior, pero valorando el hecho de considerar que las palabras consecutivas de la pregunta aparezcan en la misma frase. De este modo se consigue mejorar los resultados del sistema. Supongamos la pregunta ya citada ("vacas locas en Europa") y los documentos indicados en (28):

(28) $< DOCNO > 1$

Europa fortaleció drásticamente sus medidas para evitar la expansión del mal de las **vacas locas**.
La propuesta comunitaria incluye la prohibición de alimentar al ganado con raciones elaboradas con huesos o carne de otros animales.
Se han propuesto medidas más radicales que las utilizadas hasta la fecha.

$< DOCNO > 2$

Las amas de casa se han vuelto **locas**.
Según estadísticas realizadas en **Europa**, prefieren despojos de carne de cerdo, debido a su bajo precio que otros tipos de carnes. Esto sucede en momentos en que

el consumo de carne de **vaca** pende de un hilo.

Las apariciones de los términos de la pregunta en los documentos 1 y 2 se reflejan en las tablas 4.10 y 4.11. En estas tablas se describe en cada una de las columnas las palabras de la pregunta, en cada fila las frases que forman el documento, y en la intersección de las mismas aparece un 1 si la palabra aparece en dicha frase y un 0 en caso contrario.

En ellas se puede comprobar que el número de apariciones de los términos de la pregunta es la misma en ambos documentos. Sin embargo, en la primera de ellas, los términos de la pregunta aparecen en la misma frase, mientras que en la segunda, aparecen en frases distintas. Por su contenido se puede comprobar que el primer documento es más relevante que el segundo.

Frase	vacas	locas	europa
1	1	1	1
2	0	0	0
3	0	0	0

Tabla 4.10. Ejemplo de matriz de apariciones para documento 1

Frase	vacas	locas	europa
1	0	1	0
2	0	0	1
3	1	0	0

Tabla 4.11. Ejemplo de matriz de apariciones para documento 2

Para valorar este aspecto, se modificó la medida de forma que se incrementa en un porcentaje la puntuación que aporta un término de la pregunta, cuando alguna de las palabras que le anteceden o preceden en la pregunta (una vez eliminadas las *stop-words*) se encuentra también en la misma frase. Así a la medida $IR - n_{base}$ descrita anteriormente se añade un factor α de la siguiente forma:

$$Sim(q,d) = \sum_{t \in p \wedge q} w_{p,t} * w_{q,t} * \alpha_t \qquad (4.27)$$

siendo α_t el valor 1 para un término cuyos términos anterior o posterior en la pregunta no se encuentran en la misma frase, y otro valor superior a 1 en caso contrario.

Se ha experimentado con esta medida utilizando diferentes valores de α, distintos tipos de preguntas y en diferentes tareas como RI y BR.

Se ha podido comprobar en la experimentación realizada con esta medida, que con pequeños valores de α (1,1 y 1,2) se obtienen mejores resultados que con el modelo $IR - n_{base}$. El valor de esa mejora era más significativa en las tareas de BR que en las de RI, tal como se muestra en los capítulos 5 y 6. Fundamentalmente esto se debe a que las preguntas utilizadas en la tarea de BR son mucho más cortas y suelen contener muchos elementos relacionados. Un ejemplo de estas preguntas es "Dónde están las Montañas Rocosas?".

En general, la mejora suele ser sensible cuando las preguntas contienen términos muy relacionados como los descritos en el caso del ejemplo ("vacas locas"). Una conclusión adicional sería que si se dispusiese de un analizador sintáctico esta medida sería más eficaz, ya que se podría decidir en qué casos se ha de aplicar. Otro aspecto positivo de esta medida, es que no incrementa sensiblemente el proceso de búsqueda de los documentos relevantes.

Además, la ventaja de este modelo es que permite incrementar la puntuación de documentos que contengan términos consecutivos de la pregunta en la misma frase, pero no penaliza al documento cuando aparecen de forma separada. Como ejemplo, dada la pregunta definida en (29), este modelo otorgaría mayor valor de similitud a aquellos documentos que contuviesen las palabras "Miguel" e "Indurain" en la misma frase, pero no penalizaría a aquellos que por ejemplo sólo contuviesen la segunda de las palabras.

(29) Miguel Indurain gana el tour

La medida $IR - n_{prox}$ fue la utilizada por el sistema IR-n en las conferencias CLEF-2002 (Llopis et al., 2002g) y TREC-2002 (Vicedo et al., 2003).

4.4.2 Incorporación de técnicas de expansión de la pregunta

Como ya se ha definido en el capítulo 2, las técnicas de expansión de la pregunta permiten localizar documentos relevantes que no contienen exactamente las palabras de la pregunta.

Se ha realizado una serie de estudios sobre la incorporación de estas técnicas al sistema IR-n. Los estudios utilizados fueron dos:

1. Modelos basados en thesaurus.
2. Modelos de análisis local.

Modelos basados en thesaurus. Este modelo se aplicó al sistema IR-n incorporando a la pregunta original todos los sinónimos que proporcionaba *Wordnet* a cada uno de los nombres de la pregunta. Este modelo se probó en el CLEF-2001 (Llopis y Vicedo, 2001), empeorando notablemente los resultados obtenidos. Una vez valorados los resultados, se comprobó que al no haber realizado ningún filtro sobre los sinónimos que se añadieron, y al no utilizar ninguna técnica de desambiguación del sentido de las palabras, se habían incorporado muchas palabras que no tenían nada que ver con el sentido de la pregunta original.

Modelos de realimentación de la pregunta. Dados los resultados obtenidos por el modelo anterior, éste fue sustituido por uno de *análisis local*, basado en los trabajos de Robertson y Jones (1976); Harman (1992b); Chen et al. (2002), pero adecuándolo a la filosofía del sistema IR-n.

Estos modelos introducen en la pregunta inicialmente planteada (en adelante *pregunta original*) una serie de términos formando la *pregunta expandida*. Los términos que se añaden son aquéllos que se repiten en los documentos considerados más relevantes al

utilizar la *pregunta original*. Una vez obtenidos dichos términos, se recalculan los pesos de los términos de la *pregunta original* y de los que se incorporan a la *pregunta expandida*, de forma que tengan mayor importancia los primeros.

La figura 4.5 presenta el funcionamiento de este modelo. En primer lugar se obtiene la lista de documentos relevantes, para posteriormente generar una nueva pregunta en la que aparecen tanto los términos de la *pregunta original* como los más frecuentes de los documentos relevantes.

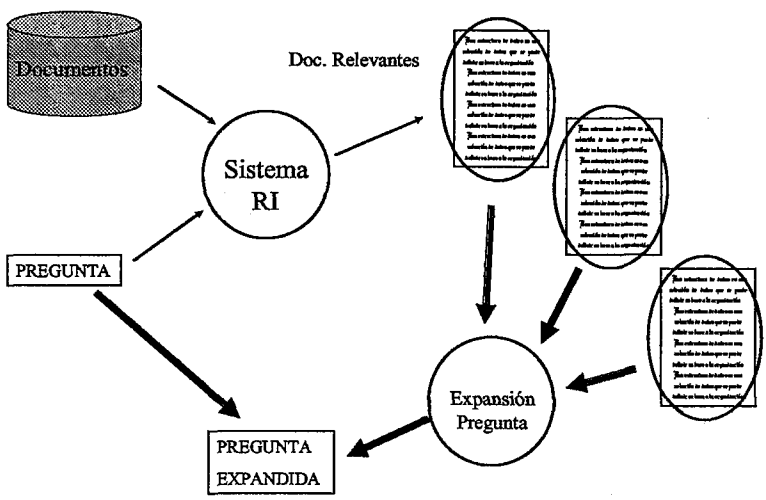


Figura 4.5. Modelo de expansión de la pregunta en sistemas de RI

La adecuación de este modelo al sistema IR-n consiste en buscar los términos a añadir en la *pregunta original*, no en los documentos más relevantes, sino en los pasajes más relevantes. Esto incrementa la relevancia de cada uno de los términos que se añaden a la pregunta original, ya que están muy cercanos a los términos de la pregunta original dentro de los documentos. Esta variación sobre el esquema inicial se muestra en la figura 4.6.

La forma en la que se ha utilizado para cada una de las preguntas ha sido la siguiente:

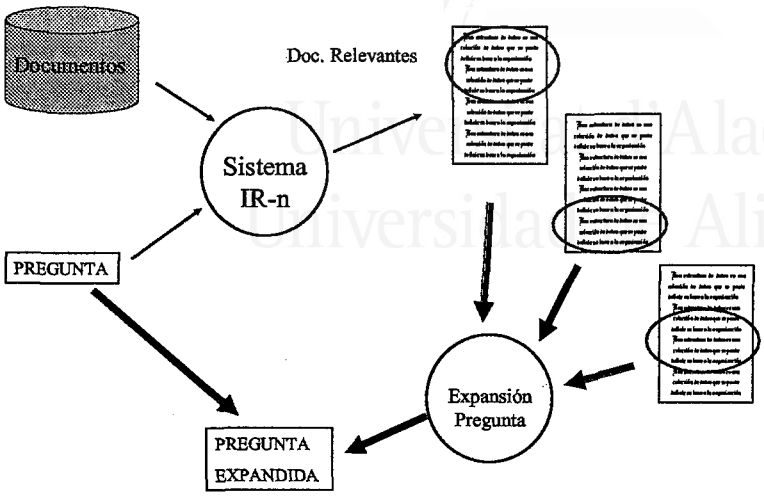


Figura 4.6. Modelo de expansión de la pregunta en el sistema IR-n

1. En primer lugar se lanza la pregunta original y se seleccionan los cinco pasajes más relevantes.
2. Se seleccionan todas las palabras, exceptuando *palabras de parada*, de todos los pasajes recuperados.
3. Se obtiene el *stem* de cada una de estas palabras, (el *stem* de cada palabra lo denominamos término).
4. Se extraen los 10 términos[3] cuyo valor de w_t sea el mayor. Este valor se calcula de la siguiente forma.

$$w_t = \frac{(m_t + 0,5) \cdot (n - n_t - m + m_t + 0,5)}{(m - m_t + 0,5) \cdot (n_t - m_t + 0,5)} \tag{4.28}$$

siendo n el número de documentos de la colección, m el número de documentos considerados como relevantes en la colección (en este caso 5 documentos[4]), n_t el número de documentos de la colección, en los que aparece el término t y m_t el número de documentos considerados relevantes, en los que aparece el término t.

[3] Este valor ha sido obtenido de forma experimental
[4] Este valor ha sido obtenido de forma experimental

Así el valor de w_t es mayor para aquellos términos que tienen un porcentaje de aparición en los primeros documentos mucho mayor que en la colección completa.

5. Estos 10 términos se añaden a la pregunta original.
6. Se modifica el factor del peso de cada término. Las normas a aplicar son las siguientes:
   - Se obtienen para los 10 nuevos términos el valor de frecuencia como el número de documentos (entre los cinco primeros pasajes) en los que aparecen, y se multiplica por 0,5. A este valor se le denomina *frecuencia de nuevo término.*
   - Si alguno de estos términos no se encuentra en la pregunta original, el valor obtenido es su *frecuencia de nuevo término.*
   - Si un término se halla en la pregunta original, se suma a la frecuencia que tenía en ésta su *frecuencia de nuevo término.*
   - Se vuelve a lanzar la pregunta con el sistema IR-n, pero utilizando los nuevos valores de frecuencia obtenidos.

Estas medidas de expansión de la pregunta fueron aplicados con éxito en la edición del CLEF-2002 (Llopis et al., 2002g) en tareas de RI, permitiendo mejorar notablemente los resultados. Esto permitió al sistema IR-n hallarse entre los cinco mejores sistemas de RI presentados sobre colecciones en español, tal y como se muestra en el capítulo 5.

Sin embargo, en experimentos realizados utilizando el mismo esquema para tareas de BR, se comprobó que los resultados eran peores que cuando no se incorporaban las técnicas de expansión de la pregunta (Llopis et al., 2002c). Estos experimentos se describen en el capítulo 6.

El estudio de los resultados obtenidos tanto en tareas de RI como de BR al utilizar técnicas de expansión de la pregunta, permiten concluir una serie de hechos:

1. La utilización de un thesaurus para localizar esos términos relacionados, sin incorporar algún proceso que los filtre (como por ejemplo técnicas de desambiguación del sentido de las palabras), no permite mejorar los resultados.
2. Los modelos de análisis local, permiten mejorar los resultados en la RI. Sin embargo, pueden provocar que documentos rele-

vantes tengan puntuaciones menores, comparándolos cuando se utilizan los términos de la pregunta original.

3. En colecciones de gran tamaño, como son las utilizadas en las pruebas de BR (colección TREC), puede ser innecesario utilizar estas técnicas de expansión, ya que para muchas preguntas existe más de un documento que contiene la respuesta, y es muy posible que al menos uno contenga varios de los términos de la pregunta original.

4.4.3 Tratamiento de las preguntas largas

En los experimentos iniciales realizados con el sistema IR-n, se detectó que funcionaba considerablemente mejor cuando las preguntas estaban formadas por pocas palabras. Así cuando se utilizaban preguntas formadas por varias frases el rendimiento del sistema se veía afectado. Esto se debía a que el sistema IR-n intenta localizar pasajes que contengan el mayor número de términos dentro de un pasaje. Las preguntas que están formadas por más de una frase intentan abarcar más de una idea con lo cual es mas difícil ubicarlas dentro de un mismo pasaje.

Por ello, se planteó un método alternativo de tratamiento de la pregunta que la descompone en una serie de subpreguntas, cada una de las cuales está formada por una de las frases que forman la pregunta originalmente planteada. Posteriormente cada subpregunta es procesada por el sistema IR-n de forma individual. Finalmente, la similitud del documento con respecto de la pregunta original se obtiene en función de las similitudes obtenidas para cada subpregunta.

Así, si la pregunta está formada por m frases según la forma:

$$Q = (f_1, f_2..f_m) \tag{4.29}$$

Los pasos que se realizan para calcular la similitud de dicha pregunta Q con respecto a un documento D son los siguientes:

1. Se generan tantas subpreguntas q_i como frases tiene la pregunta inicial Q, de forma que cada pregunta esta formada por una frase de esta.

$$q_i = f_i, : i \in [1..m] \tag{4.30}$$

2. Se calcula la similitud X_i de cada una de las subpreguntas con respecto del documento utilizando el sistema IR-n.

$$sim(q_i, D) = X_i \tag{4.31}$$

3. La similitud del documento se asigna en función de la mejor similitud obtenida por cada una de las subpreguntas q_i.

$$sim(Q, D) = \sum_{\forall_{i \in [1..n]}} sim(q_i, D) \tag{4.32}$$

Un ejemplo de la descomposición de una pregunta sería el descrito en (30)

(30) Pregunta original

Relevant documents give information on the discovery of pesticides in baby food. They report on different brands, supermarkets, and companies selling baby food which contains pesticides. They also discuss measures against the contamination of baby food by pesticides.

Preguntas a procesar

Pregunta 1

Relevant documents give information on the discovery of pesticides in baby food.

Pregunta 2

They report on different brands, supermarkets, and companies selling baby food which contains pesticides.

Pregunta 3

They also discuss measures against the contamination of baby food by pesticides.

El hecho de asignar a cada documento la suma de similitudes obtenidas con las diferentes subpreguntas, se obtuvo de forma experimental, comparándose esta posibilidad con la de asignar a cada documento la mejor similitud. Estos experimentos se describen en Llopis et al. (2002e).

El hecho de que definir la similitud sobre la pregunta original en base a la suma de similitudes de cada subpregunta obtenga mejores resultados que la de asignar la mayor similitud, se debe a

que al dividir las preguntas en subpreguntas (cada una formada por una frase), puede ocurrir que algunas de ellas estén formadas por un mayor número de palabras o por palabras más discriminantes que las contenidas en otras subpreguntas. Este hecho provoca que las puntuaciones obtenidas por los documentos para una frase en general sean sensiblemente mayores que los obtenidos en otra. Esto puede producir que documentos algo relevantes a una subpregunta, sean considerados más relevantes que documentos muy relevantes a otra subpregunta. Esto no ocurre de la misma forma al utilizar la suma de relevancia de cada subpregunta.

Los experimentos descritos en el capítulo 5, así como los resultados obtenidos en la participación en el CLEF-2002 (Llopis et al., 2002g) demuestran que este modelo permite mejorar los resultados en el tratamiento de preguntas formadas por más de una frase.

4.5 Conclusiones

En este capítulo se ha presentado una de las partes principales del trabajo realizado, el cual ha sido la definición de un modelo de RP, denominado IR-n y que utiliza las frases como unidad de definición de los pasajes.

Este modelo, además de las ventajas propias de los modelos de RP, tiene una serie de ventajas adicionales:

1. Permite definir de forma sencilla, precisa y eficiente una serie de pasajes con estructura sintáctica completa, ya que se basa en la división de pasajes en función de las frases del documento, incorporando el solapamiento de las mismas para conseguir mayor fiabilidad.
2. Permite incorporar a las medidas de similitud básicas, la consideración de que términos de la pregunta aparecen en la misma frase. Es decir, incrementa las posibilidades de utilizar conceptos de cercanía en las medidas de similitud.
3. Es un modelo que se puede aplicar a cualquier tipo de documento aunque no esté estructurado e incluso si es de distintos

Volver al índice/Tornar a l'índex

estilos, ya que resulta sencilla la detección de los límites de las frases.

4. Es un modelo flexible al permitir adaptarse a distintos modelos de pregunta, bien en función de su tamaño o su tipo.
5. Es un modelo que permite incrementar la relación entre los términos que se añaden a la pregunta original en las técnicas de expansión de la pregunta.
6. Se ha definido en el sistema una forma de procesar las preguntas largas (formadas por más de una frase), para que no se vea afectado el rendimiento del sistema.
7. Gracias a la forma en la que se ha implementado, se permite realizar de forma eficiente los procesos de indexación y búsqueda.
8. Es un modelo que permite adaptarse de forma eficaz a diferentes tipos de colecciones de documentos y de estilos de preguntas.
9. Mejora los resultados de otros sistemas de RI, tal como se demostrará en el capítulo 5.
10. El modelo definido permite su integración en otras tareas como son la BR y la SID, que se analizarán en el capítulo 6.
11. El sistema IR-n se puede utilizar con éxito en diferentes idiomas, con sólo incluir las *listas de parada* y proceso de *stemming* de cada idioma.

Una vez descrito el sistema IR-n, en los capítulos 5 y 6 se mostrará la experimentación realizada con el sistema IR-n en diferentes tareas (RI, BR y SID) y utilizando diferentes colecciones. Esta experimentación tiene como objetivo fijar las características parametrizables del sistema IR-n descritas en este capítulo. Posteriormente el sistema IR-n se ha evaluado de forma externa mediante su participación en congresos que evalúan este tipo de sistemas (TREC y CLEF).

5. Evaluación del sistema IR-n en tareas de Recuperación de Información

En este capítulo se presenta la evaluación en tareas de RI del sistema IR-n.

En primer lugar se describe la problemática de las tareas de evaluación en este tipo de sistemas, se presentan las diversas propuestas existentes, y se justifica la elección del método de evaluación utilizado en este trabajo.

En segundo lugar se detalla el proceso de entrenamiento del sistema. Se describen las características de las colecciones de test utilizadas, las tareas a realizar, así como el desarrollo de los experimentos realizados en el proceso de entrenamiento del sistema IR-n. El objetivo de estos experimentos es determinar las características del modelo de recuperación propuesto que permitan obtener los mejores resultados.

En tercer lugar se detalla el proceso de evaluación al que se ha sometido al sistema. Esta evaluación se ha realizado utilizando colecciones de test diferentes a las empleadas en el proceso de entrenamiento, justificándose los motivos de esta elección.

Finalmente se analizan los resultados obtenidos en esta evaluación y se comparan frente a los modelos de RI y RP más conocidos.

5.1 La evaluación de los sistemas de recuperación de información

La evaluación de un sistema, es un elemento fundamental para determinar la calidad del mismo. La evaluación no sólo se debe restringir a la medición de la eficacia de un sistema, sino que también debe contemplar la eficiencia y facilidad de interacción con los usuarios que proporciona.

La eficacia de un sistema de RI se valora desde el punto de vista estadístico, a nivel del número de documentos relevantes recuperados.

La eficiencia valora aspectos relacionados con la complejidad temporal y espacial del sistema, es decir, un sistema de RI será mejor cuando tanto el tiempo de respuesta como la cantidad de recursos utilizados sean los mínimos posibles. También, un sistema puede valorarse desde el punto de vista del usuario, a nivel de la satisfacción que produce su utilización, o de las facilidades que éste proporciona en su interacción con él.

Fundamentalmente, la evaluación que se va a seguir en este capítulo se centra en la valoración de la eficacia del sistema, esto es, intentar valorarlo en función de la relevancia o no de los documentos que recupera ante una determinada pregunta sobre una colección de documentos. No obstante, no se va a olvidar la valoración de a eficiencia del sistema, ya que se han realizado pruebas que la analizan. Además, en el capítulo siguiente, dentro de las tareas de *Selección Interactiva de Documentos* (en adelante SID), se valorarán las ventajas que pueda aportar el sistema IR-n en la interacción con los usuarios.

En las siguientes subsecciones se presentan las medidas de evaluación que se van a utilizar para poder contrastar los diferentes sistemas, y se describen las ventajas que aporta el uso de colecciones de test dentro de la tarea de evaluación.

5.1.1 Medidas de evaluación

Las medidas de evaluación de los sistemas de RI más utilizadas son la de *precisión* y la *cobertura*. La cobertura (o *recall*) mide la habilidad del sistema para recuperar todos los documentos relevantes que existen en la colección. Se calcula de la siguiente forma:

$$Cobertura = \frac{NRR}{NRC} \tag{5.1}$$

siendo:
NRR = Número de documentos relevantes recuperados

NRC = Número de documentos relevantes en la colección

La precisión (o *precision*), mide la habilidad del sistema para recuperar sólo aquellos documentos que son relevantes. Se calcula de la siguiente forma:

$$Precision = \frac{NRR}{NDR} \tag{5.2}$$

siendo:

NRR = Número de documentos relevantes recuperados
NDR = Número de documentos recuperados

Tanto la cobertura como la precisión son medidas de conjunto, es decir, evalúan la calidad de un conjunto desordenado de documentos relevantes recuperados. Por ejemplo, supongamos que se dispone de una colección de 1000 documentos, la cual contiene 50 documentos relevantes para una pregunta determinada. Si el sistema de RI recupera 100 documentos de los cuales 40 son relevantes, el sistema ha obtenido una cobertura de un 0,8 ó 80% (40 documentos relevantes recuperados sobre 50 relevantes que existen en la colección) y una precisión de un 0,4 ó 40% (de los 100 documentos recuperados hay 40 que son relevantes).

5.1.2 Colecciones de test

Supongamos que disponemos de un sistema de RI que recupera un conjunto de documentos considerados relevantes para una determinada pregunta o tema. Para poder calcular las medidas de cobertura y precisión obtenidas por dicho sistema, se debe conocer tanto la relevancia de los documentos que recupera como la relevancia del resto de los documentos que forman la colección.

Por ello, la evaluación de los sistemas de RI puede ser una tarea prácticamente imposible de conseguir si se realiza de forma totalmente manual. Por ejemplo, el verificar el rendimiento de un sistema de RI que devuelve 200 documentos entre una colección de 100.000, supondría estudiar la relevancia de los 200 documentos

recuperados para dicha pregunta, y además verificar si existen más documentos relevantes entre los 99.800 no recuperados.

Para solucionar este problema, la propuesta que mayor éxito ha tenido hasta el momento ha sido la de utilización de colecciones de test. Una colección de test está formada fundamentalmente por tres elementos: una colección de documentos, una colección de preguntas y los juicios de relevancia que indican qué documentos de la colección son relevantes para cada una de las preguntas.

La mayor ventaja que aportan las colecciones de test es que permiten automatizar el proceso de cálculo del rendimiento de los sistemas de RI utilizando las preguntas de la colección de test. Además permiten comparar diferentes sistemas de RI utilizados sobre la misma colección.

Existen dos formas de generar una colección de test: según la *tradición Cranfield* o a través del *pooling*.

La tradición Cranfield (Jones, 1981) consiste en disponer de una colección de documentos, así como los criterios de relevancia completos, es decir, se conoce para cada pregunta si cada uno de los documentos que forman la colección es relevante o no. Esto tiene como principal ventaja que la cobertura de cualquier pregunta se puede conocer, ya que se dispone del número total de documentos relevantes para cada una de las preguntas existentes en la colección. Los principales inconvenientes de esta aproximación son varios. En primer lugar el coste necesario para definir los juicios de relevancia, ya que se debe estudiar la colección completa para cada pregunta. En segundo lugar, el hecho de que este modelo presupone que los criterios de relevancia son fijos en el tiempo. Un ejemplo de este modelo de evaluación de un sistema de RI es la colección Cranfield, formada por 1.400 documentos (aproximadamente 1,5 Mb de tamaño) y 22 preguntas.

Otro enfoque para la construcción de colecciones de test es el denominado *pooling*. Éste es el modelo utilizado en las conferencias TREC y CLEF. Este enfoque consiste en recuperar los documentos relevantes para cada pregunta utilizando diferentes sistemas. Esto se suele realizar a través de diferentes concursos en los que cada participante indica los documentos considerados más relevantes para cada pregunta según su sistema de RI. A conti-

nuación, la organización juzga la relevancia de los n (usualmente 1000) primeros documentos que cada participante ha considerado más relevantes para cada pregunta. El conjunto de documentos relevantes para cada pregunta se forma con la unión de todos los conjuntos de documentos realmente relevantes obtenidos por cada participante. Este enfoque tiene como principal ventaja que disminuye el número de documentos para los que se debe estudiar su relevancia, ya que no es necesario determinar para cada pregunta la relevancia de todos los documentos de la colección, sino únicamente de aquellos documentos que han sido considerados relevantes por al menos uno de los participantes. El inconveniente principal de esta aproximación, es que la cobertura real no se conoce, ya que puede haber documentos relevantes para una pregunta que no hayan sido recuperados por ninguno de los sistemas utilizados.

La elección de la colección de test a utilizar en procesos de evaluación de sistemas de RI es de gran importancia. Dicha importancia reside en varios aspectos:

- *El tamaño de la colección.* Cuanto mayor sea la base de datos documental y el número de preguntas a evaluar, mejor se ajustará a la realidad la medida resultante del comportamiento del sistema.
- *La calidad de la colección de preguntas de test.* Esta calidad depende de la variedad de tipos de preguntas realizadas, de la diversidad de construcciones utilizadas y sobre todo, de si esas preguntas corresponden o no a requerimientos "reales" de información.
- *La colección de documentos.* En este caso, la calidad depende principalmente de la variedad de documentos de la colección y de que éstos sean documentos originales sin ningún tipo de tratamiento especial.

Las colecciones utilizadas en las conferencias CLEF cumplen estos requisitos y por ello se han seleccionado para realizar el proceso de entrenamiento y evaluación del sistema IR-n presentado en esta tesis. Además, para garantizar que los estudios que se

describen a continuación no son dependientes de una colección en concreto, se han seleccionado dos de estas colecciones para realizar el proceso de entrenamiento del sistema.

A continuación se detallan las especificaciones de la tarea de RI desarrollada en el CLEF-2001 y las características de la colección de test generada a partir de las pruebas realizadas en dicha convocatoria.

5.2 Descripción de la tarea de RI monolingüe CLEF-2001

El objetivo de las conferencias CLEF es evaluar el rendimiento de diferentes sistemas en tareas de RI monolingüe y multilingüe. Realmente la tarea principal del CLEF es ésta última, no obstante, las experiencias obtenidas en las diversas conferencias CLEF realizadas, han demostrado lo positiva que es la experiencia en la tarea de RI monolingüe como primer paso a la realización de un sistema de RI multilingüe.

El objeto de este apartado es describir el funcionamiento de la tarea de RI monolingüe en las conferencias CLEF.

5.2.1 Especificación de la tarea

Los participantes reciben una colección de documentos y un conjunto de preguntas para cada una de las tareas que vayan a desarrollar.

Para cada una de estas preguntas, el participante ha de obtener una lista de documentos ordenados en función de la relevancia a cada una de ellas. El número máximo de documentos que se pueden devolver para cada pregunta es de 1000. Posteriormente, la organización del congreso determina la relevancia de los documentos presentados y evalúa cada sistema en función de una serie de medidas. Así, los elementos principales de este proceso son los siguientes:

1. La colección de documentos
2. La colección de preguntas

3. Ficheros de resultados
4. Criterios de relevancia
5. Medidas de evaluación

Estos elementos se describen detalladamente a continuación en las siguientes subsecciones.

5.2.2 Colecciones de documentos

Las conferencias CLEF disponen de colecciones de documentos para cada uno de los idiomas con los que se realizan las pruebas. En este trabajo han sido objeto de estudio dos de ellas, una en inglés y la otra en español. Las colecciones son las siguientes:

- **Los Angeles Times.** Esta colección está formada por un conjunto de artículos publicados en el periódico Los Angeles Times en el periodo comprendido desde el 1 de enero al 31 de diciembre de 1994. La colección se halla escrita en inglés. En adelante las referencias a esta colección se realizarán empleando el término LATimes.
- **Noticias de agencia EFE.** Esta colección esta formada por un conjunto de noticias de la agencia EFE realizadas del 1 de enero al 31 de diciembre de 1994. La colección se halla escrita en español. En adelante las referencias a esta colección se realizarán empleando el término EFE.

La tabla 5.1 muestra las principales características de estas colecciones referentes a su tamaño (en bytes), número de documentos y la media de términos por documento. Como se puede observar en dicha tabla, ambas colecciones tienen un tamaño similar, si bien, la colección EFE contiene casi el doble de documentos que la LATimes. Esto se debe a que los documentos de la colección EFE son considerablemente más pequeños, en cuanto a número de palabras, que los de la colección LATimes.

En la tabla 5.2 se puede ver información más específica de la colección referida al tamaño de los documentos medidos en frases. En esta tabla se muestra el mayor y menor número de frases que contiene algún documento de la colección, así como la media y

Colección	Idioma	Tamaño (Mb)	Num. Docs.	Media Term/Doc
LATimes	Inglés	425	113.005	167,33
EFE	Español	509	215.738	120,24

Tabla 5.1. Información general de las colecciones de documentos en inglés y español del CLEF

desviación típica referente al número de frases que contienen todos los documentos de la colección.

Colección	Min	Max	Media Frases/Doc	Desviación Típica
LATimes	1	1233	31,51	22,24
EFE	2	225	11,56	4,30

Tabla 5.2. Información sobre frases en las colecciones de documentos en inglés y español del CLEF

Como se puede observar, el número medio de frases que constituyen un documento en la colección LATimes es casi el triple que en la colección EFE. Además la dispersión en cuanto al tamaño de documentos medido en frases es mucho mayor en la colección LATimes que en la colección EFE.

Los gráficos 5.1 y 5.2 muestran la distribución de los documentos de las colecciones en función del número de frases que los forman. En el eje de las ordenadas se indica el número de frases por intervalos de 5 unidades. En el eje de las abscisas se indican el número de documentos que tienen un número de frases comprendidas en dicho intervalo.

Observando los gráficos se puede ver que en la colección EFE, la mayoría de documentos tienen un tamaño similar (cercano a las 11 frases), mientras que la colección LATimes está formada por documentos de tamaño muy heterogéneo.

5.2.3 Colecciones de preguntas

Las colecciones de preguntas utilizadas en cada edición en el CLEF son las mismas para todas las tareas de RI monolingüe, aunque evidentemente cada una de ellas está escrita en el idioma definido en la tarea.

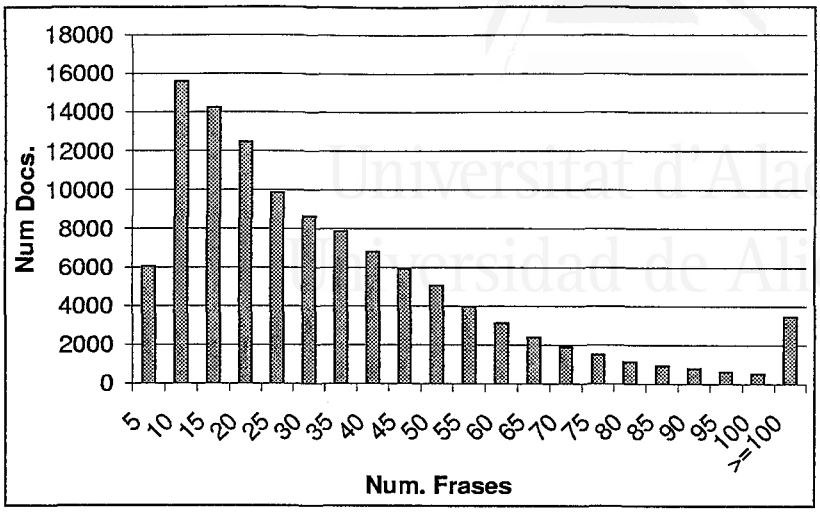


Figura 5.1. Cantidad de documentos según número de frases que lo forman. Colección LATimes

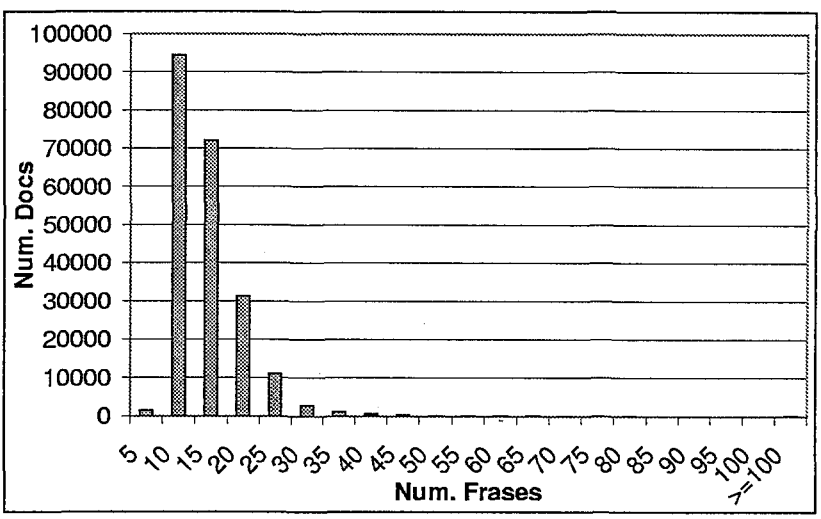


Figura 5.2. Cantidad de documentos según número de frases que lo forman. Colección EFE

El formato de las preguntas utilizadas en el CLEF se define en base a tres apartados: *título*, *descripción* y *narrativa*.

El título contiene las palabras clave de la búsqueda y suele estar formado por entre 2 y 4 palabras. La descripción es similar al título, y describe de forma escueta en una sola frase la tarea a realizar. La narrativa detalla en mayor medida el objeto de la pregunta, y suele estar formada por varias frases.

En (31) se muestra una de las preguntas de esta colección en español. En este ejemplo se observa que la parte de descripción de la pregunta solicita la búsqueda de documentos que den información sobre pesticidas hallados en alimentos infantiles, mientras que la parte narrativa da muchos más detalles de la búsqueda, solicitando información de supermercados y compañías que han vendido alimentos infantiles con pesticidas, así como información sobre las medidas para evitar esta contaminación de los alimentos infantiles.

(31) **Título:** Pesticidas en alimentos para bebes.
Descripción: Encontrar noticias sobre pesticidas en alimentos para bebes.
Narrativa: Los documentos relevantes proporcionan información sobre el descubrimiento de pesticidas en alimentos para bebes. Se informa sobre diferentes marcas, supermercados y compañías que ofrecieron alimentos para bebes que contenían pesticidas. Se discuten también medidas contra la contaminación de alimentos para bebes con pesticidas.

Dentro de las pruebas realizadas en esta tesis, y con el objeto de uniformizar el tratamiento de todas las preguntas, se han considerado dos tipos de las mismas: cortas y largas. Las primeras están formadas por los campos título y descripción, y las segundas, por los tres campos: título, descripción y narrativa.

Dada la pregunta descrita en (31), la pregunta corta y la larga generadas a partir de la misma se muestran en (32) y (33).

(32) **PREGUNTA CORTA**

Pesticidas en alimentos para bebes.
Encontrar noticias sobre pesticidas en alimentos para bebes.

(33) **PREGUNTA LARGA**

Pesticidas en alimentos para bebes.
Encontrar noticias sobre pesticidas en alimentos para bebes.
Los documentos relevantes proporcionan información sobre el descubrimiento de pesticidas en alimentos para bebes. Se informa sobre diferentes marcas, supermercados y compañías que ofrecieron alimentos para bebes que contenían pesticidas. Se discuten también medidas contra la contaminación de alimentos para bebes con pesticidas.

En la edición del CLEF-2001 se utilizaron juegos formados por cincuenta preguntas para cada idioma. Las características generales de los juegos de preguntas se muestran en la tabla 5.3. Como puede observarse no todas las preguntas planteadas tienen necesariamente algún documento relevante en la colección, o al menos no ha sido encontrado por alguno de los participantes. Es por esto por lo que aunque los juegos originales contienen 50 preguntas, sólo serán utilizadas las 47 y 49 preguntas que tienen algún documento relevante en las colecciones de LATimes y EFE respectivamente.

Idioma	Edición	Preguntas Num.	Número de preguntas con respuesta
Inglés	CLEF-2001	41-90	47
Español	CLEF-2001	41-90	49

Tabla 5.3. Colecciones de preguntas en inglés y castellano del CLEF utilizadas en la experimentación

5.2.4 Ficheros de resultados

Una vez efectuados los procesos de tratamiento de las colecciones de documentos y preguntas, así como el proceso de obtención de documentos relevantes para cada una de ellas, los participantes deben entregar un fichero a la organización con el formato indicado en (34).

(34)
```
41 Q0 EFE19940407-03243 0 31.3664 1
41 Q0 EFE19940406-02595 1 30.3146 1
41 Q0 EFE19940405-01875 2 27.3849 1
41 Q0 EFE19940406-02364 3 24.9771 1
41 Q0 EFE19940519-11553 4 21.9647 1
```

Donde cada uno de los datos (separado por blancos) se utiliza de la siguiente forma:

1. Número de la pregunta.
2. Campo no utilizado.
3. Número de documento recuperado. Es el contenido de la etiqueta <DOCNO> del documento considerado relevante.
4. Posición. Debe contener valores entre 0 y n, indicando 0 la posición del documento más relevante y n la del menos relevante de los seleccionados. En las últimas conferencias CLEF n ha tomado el valor 999.
5. Similitud otorgada por el sistema de RI al documento. Cualquier documento que ocupa una posición superior a otro siem-

pre deberá tener un valor de similitud superior o igual a la de este último.

6. Identificador de la prueba. Permite diferenciar diferentes pruebas realizadas por el mismo sistema.

5.2.5 Criterios de relevancia

Para cada una de las preguntas se genera un fichero que contiene la unión, sin duplicados, de todos los documentos considerados relevantes por los diversos participantes. De forma manual se evalúa la relevancia de cada documento a dicha pregunta y se genera un fichero con el formato definido en (35).

(35)
```
41 0 LA010194-0002 0
41 0 LA010194-0004 0
41 0 LA010194-0033 0
41 0 LA010694-0009 1
42 0 LA091594-0012 0
42 0 LA091594-0102 0
42 0 LA091594-0104 1
42 0 LA091594-0125 0
42 0 LA091594-0126 0
42 0 LA091594-0127 0
```

Cada uno de estos campos se utiliza de la siguiente forma:

1. Número de la pregunta.
2. Campo no utilizado.
3. Nombre del documento recuperado.
4. Valor que indica si el documento es relevante o no para dicha pregunta. Tomará el valor uno en el primer caso y cero en el segundo.

Como ya se ha indicado en apartados anteriores, la información disponible sobre la relevancia no es completa, ya que si en pruebas

posteriores se considera como relevante un documento que no está incluido en este fichero, a todos los efectos se consideraría como no relevante, aunque realmente no se dispone de información real para juzgar dicha relevancia.

La organización dispone de un programa denominado "trec_eval" que tiene como entrada los ficheros de resultados obtenidos por el participante y los ficheros de relevancia generados por la organización. Con esta información, el programa obtiene de forma automática los resultados de evaluación del sistema. Este programa de evaluación se ha utilizado para evaluar de forma automática el rendimiento del sistema IR-n en todas las pruebas realizadas.

5.2.6 El proceso de evaluación

Una vez los participantes han enviado sus listas de documentos relevantes, la organización evalúa la relevancia o no de todos los documentos presentados, y obtiene el rendimiento final de los sistemas que han participado en la prueba. Esto le permite al participante conocer, por una parte, los resultados que ha obtenido su sistema y por otra, una comparativa general, pregunta por pregunta, de su sistema frente al resto de sistemas participantes. Estos resultados se muestran en una serie de tablas y gráficas que facilitan la interpretación de los resultados.

Estos resultados se basan fundamentalmente en la valoración de las medidas comentadas antes: precisión y cobertura. No obstante, se tienen en cuenta otros factores, ya que, a pesar de que los valores de cobertura y precisión permiten medir la eficacia de un sistema de RI, tienen como principal inconveniente el que se limitan a valorar cantidades totales de documentos relevantes recuperados, pero no consideran el orden en el que han sido recuperados. Es lógico pensar que dados dos sistemas de RI que recuperan el mismo número de documentos relevantes, será preferible aquel sistema que recupere dichos documentos relevantes en primer lugar. Por ello, en las conferencias CLEF no sólo se limitan a valorar esa precisión y cobertura total obtenida por los sistemas de RI, sino que utilizan una serie de tablas y medidas

que permiten tener en cuenta también la posición en la que se recuperan los documentos relevantes.

Así, los resultados de la evaluación de cada uno de los sistemas se presentan fundamentalmente en cuatro tablas de resultados y una gráfica:

- Tabla de estadísticas generales.
- Tabla de medias de niveles de precisión y cobertura.
- Tabla de medias de precisión para cada pregunta.
- Tabla de medias de niveles de documentos recuperados.
- Gráfico de cobertura-precisión.

Estas medidas se describen individualmente en los siguientes apartados.

5.2.7 Tabla de estadísticas generales

Esta tabla ofrece una serie de información global de toda la prueba y permite obtener la precisión y la cobertura total obtenida por el sistema. Contiene la siguiente información:

- Número de preguntas realizadas.
- Número total de documentos recuperados para todas las preguntas.
- Número total de documentos relevantes en la colección para todas las preguntas.
- Número total de documentos relevantes recuperados.

Queryid (Num): 47
Total number of documents over all queries :
Retrieved: 47000
Relevant: 856
Rel_ret: 796

Tabla 5.4. Tabla de estadísticas generales

El ejemplo de la tabla 5.4 muestra cómo el sistema evaluado ha recuperado 47.000 documentos para las 47 preguntas (1.000 por pregunta) de los cuales 796 documentos eran relevantes. Además

se especifica que en la colección había un total de 856 documentos relevantes para las 47 preguntas.

Con la información contenida en esta tabla se pueden obtener los valores de precisión y cobertura tal como se ha comentado en apartados anteriores. Este sistema hubiese obtenido una precisión del 1,69 % (856/47.000) y una cobertura del 93%(796/856).

5.2.8 Tabla de medias de niveles de precisión y cobertura

Esta tabla muestra dos informaciones:

1. *La precisión que obtiene el sistema a once valores de cobertura determinados (del 0 al 1 con incrementos de 0,1).* Estos valores permiten comparar el rendimiento de diferentes sistemas y dibujar el gráfico de cobertura-precisión, descrito en la sección 5.2.11. Cada valor de la media de precisión a un nivel de cobertura determinado se calcula sumando las precisiones obtenidas en dicho nivel de cobertura para cada pregunta y dividiendo por el número de preguntas. La precisión que se utiliza en esta medida es la precisión interpolada. La precisión interpolada a un nivel de cobertura R^i se define como la precisión máxima a todos los puntos p tales que se cumpla:

$$R^{i-1} \leq p \leq R^i \tag{5.3}$$

2. *Media de precisión no interpolada*
   Éste es un valor que refleja el rendimiento sobre todos los documentos relevantes, no solo considerando el porcentaje de documentos relevantes recuperados, sino además, valorando el orden en el que han sido recuperados. La medida recompensa los sistemas que recuperan los documentos relevantes en los primeros lugares.
   Esta medida no es la media de la precisión a todos los niveles de cobertura. En su lugar, es la media de la precisión obtenida después que cada documento relevante es recuperado. La forma de calcularla es la siguiente:
   - Cuando no se ha recuperado todavía ningún documento relevante, la precisión es 0.

- Cada vez que se obtiene un documento relevante, se calcula la precisión.

Por ejemplo, si una pregunta tiene 4 documentos relevantes, los cuales se recuperan en las posiciones 1, 2, 4 y 7, se calcula la precisión una vez llegados esos niveles. O sea, la precisión toma los siguientes valores:

- 1 en el primer documento relevante recuperado (1 documento relevante/1 documento recuperado)
- 1 en el segundo documento relevante recuperado (2 documentos relevantes/2 documentos recuperados)
- 0,75 en el tercer documento relevante recuperado (3 documentos relevantes/4 documentos recuperados)
- 0,57 en el cuarto documento relevante recuperado (4 documentos relevantes/4 documentos recuperados)

La media de estos cuatro valores es 0,83.

La tabla 5.5 muestra un ejemplo de este modelo de tablas. En ella se puede apreciar las medias de precisiones interpoladas a los 11 niveles de cobertura fijados (por ejemplo, el sistema ha obtenido una precisión del 57,65 % cuando la cobertura obtenida es del 20%) y la media de precisión no interpolada (en este caso es del 41,89%).

Interpolated Recall - Precision Averages:
at 0.00 0.6972
at 0.10 0.6234
at 0.20 0.5765
at 0.30 0.5240
at 0.40 0.4785
at 0.50 0.4330
at 0.60 0.3667
at 0.70 0.3271
at 0.80 0.3036
at 0.90 0.2711
at 1.00 0.2090
Average precision (non-interpolated) for all rel docs(averaged over queries)
0.4189

Tabla 5.5. Tabla de medias de precisión y cobertura interpoladas

5.2.9 Tabla de medias de precisión por pregunta

Esta tabla contiene para cada una de las preguntas, el valor de la media de precisión, calculada de la misma forma que la de la tabla anterior.

Average precision individual queries)
Query 41: 1.0000
Query 42: 0.3919
Query 43: 1.0000
Query 44: 0.2817

Tabla 5.6. Tabla de medias de precisión por pregunta

En la tabla 5.6 se pueden ver las precisiones que se han obtenido en las preguntas 41, 42, 43 y 44. El principal objetivo de esta medida es permitir valorar en qué preguntas se han obtenido mejores y peores resultados.

5.2.10 Tabla de medias de niveles de documentos recuperados

Esta tabla contiene los valores de la precisión obtenidos una vez se ha recuperado un número determinado de documentos. Los valores de corte que se utilizan son los de 5, 10, 15, 20, 30, 100, 200 y 1.000 documentos. Cuando hay varias preguntas, la precisión a un nivel de corte es la media de los valores de precisión para cada una de las preguntas en esos niveles de corte.

Esta tabla también contiene el valor de la *precisión-R*. Este valor muestra la precisión a los R documentos recuperados, siendo R el número de documentos relevantes para una pregunta. El valor medio de la precisión-R se calcula tomando la media de los valores individuales de precisión-R para cada pregunta.

Por ejemplo, dada una prueba que consiste en 2 preguntas, la primera de ellas con 50 documentos relevantes y la segunda con 10. Si el sistema recupera 17 documentos relevantes en los primeros 50 documentos devueltos para la primera pregunta, y 7 en los 10 primeros devueltos para la segunda, el valor de la precisión-R es ((17/50)+(7/10))/2, o lo que es lo mismo 0,52.

Precision:
At 5 docs: 0.4000
At 10 docs: 0.3277
At 15 docs: 0.3078
At 20 docs: 0.2745
At 30 docs: 0.2326
At 100 docs: 0.1198
At 200 docs: 0.0714
At 500 docs: 0.0324
At 1000 docs: 0.0169
R-Precision (precision after R (= num_rel for a query) docs retrieved):
Exact: 0.4050

Tabla 5.7. Tabla de precisión media no interpolada

En la tabla 5.7 se puede ver que el sistema evaluado obtiene una precisión media del 40% a los 5 documentos recuperados y una precisión-R del 40,5%.

5.2.11 Gráfico de cobertura-precisión

Este gráfico se calcula en función de los 11 valores de corte utilizados en la tabla de medias de precisión interpolada por nivel de cobertura. Así, en este gráfico se muestran los valores de cobertura a diferentes niveles de precisión y viceversa. En el eje de las ordenadas se indican los valores de cobertura y en el de las abscisas los de precisión.

Normalmente este gráfico es una línea descendente de izquierda a derecha, dado que a medida que se incremente el número de documentos relevantes recuperados (se incrementa la cobertura), más documentos no relevantes se recuperan (la precisión disminuye).

Este gráfico es el más común para comparar sistemas. Los gráficos de dos sistemas se pueden superponer en el mismo gráfico determinando cuál de los dos es mejor. En el momento de comparar se suelen medir en 3 rangos de cobertura, 0 a 0,2, 0,2 a 0,8 y 0,8 a 1. Esos rangos caracterizan la alta precisión, cobertura media y alta cobertura respectivamente.

En la figura 5.3 se puede ver el gráfico de los valores de cobertura y precisión correspondientes al ejemplo descrito en la tabla 5.5.

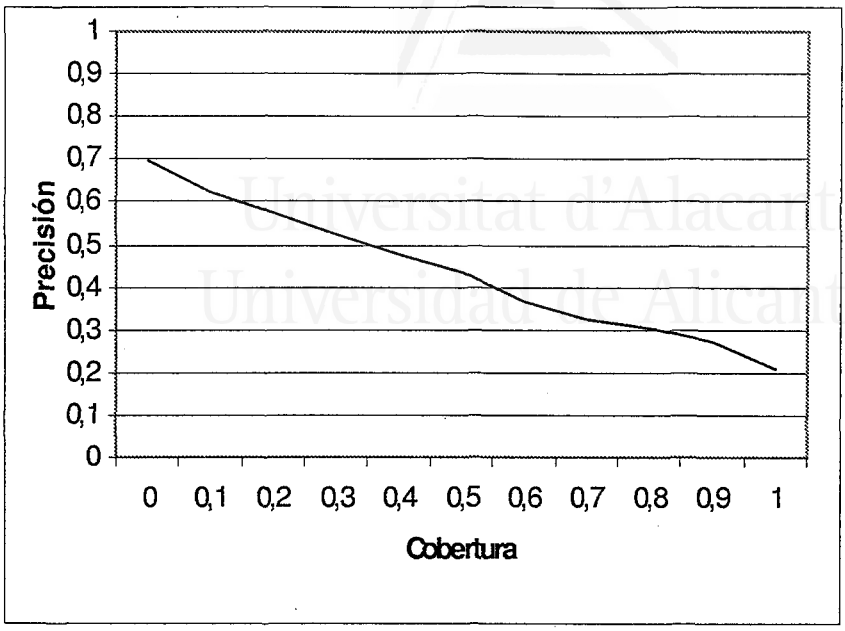


Figura 5.3. Gráfico de Cobertura y Precisión

5.3 Entrenamiento del sistema IR-n

El proceso de entrenamiento pretende obtener los valores de los parámetros de los que depende el sistema IR-n con el objetivo de maximizar su rendimiento en tareas de RI. Los parámetros a ajustar son:

1. Tamaño de pasaje.
2. Grado de solapamiento a utilizar en la definición de pasajes.
3. Medida de similitud a utilizar (de las dos medidas de similitud definidas en el capítulo 4: $IR - n_{base}$ e $IR - n_{prox}$).
4. Forma de procesar las preguntas largas.
5. Forma de incorporar al sistema IR-n el módulo de expansión de la pregunta.

Con el objeto de no basarse únicamente en una colección de test, el sistema IR-n se ha entrenado de forma individual con dos colecciones de test de características bien diferenciadas que han sido descritas en las secciones anteriores: LATimes (colección en

inglés de documentos de tamaño heterogéneo) y EFE (colección en español de documentos de tamaño similar). Ambas colecciones fueron empleadas en la edición CLEF del año 2001.

A continuación se describen cada uno de los experimentos realizados.

5.3.1 Experimentos realizados

- **Experimento 1. Definición del sistema base.** Este experimento tiene como objetivo el cálculo de unos resultados base que permitan comparar los resultados obtenidos por el sistema IR-n con la misma colección de test.
- **Experimento 2. Obtención del tamaño óptimo de los pasajes.** Este experimento tiene como objetivo valorar la medida $IR - n_{base}$, descrita en la subsección 4.2.2, y determinar el tamaño de pasaje a utilizar con el que se obtienen los mejores resultados para cada una de las colecciones (LATimes y EFE) empleando los dos tipos de preguntas propuestos (cortas y largas).
- **Experimento 3. Obtención del grado de solapamiento óptimo.** Este experimento tiene como objetivo ajustar el grado de solapamiento utilizado en el sistema IR-n, para estudiar cómo afecta al rendimiento y eficiencia del sistema.
- **Experimento 4. Aplicación de medidas de proximidad.** Este experimento tiene como objetivo estudiar la influencia en los resultados obtenidos de la aparición de palabras consecutivas de la pregunta en la misma frase de un documento. En estas pruebas se utilizará la medida $IR - n_{prox}$, descrita en la subsección 4.4.1.
- **Experimento 5. Valoración de separación de preguntas largas.** Este experimento tiene como objetivo valorar los resultados obtenidos al procesar las preguntas largas tratadas como una serie de preguntas cortas. Este modelo de tratamiento se definió en la subsección 4.4.3.
- **Experimento 6. Expansión de la pregunta.** Este experimento tiene como objetivo analiza la influencia de las técnicas de expansión de la pregunta, estudiadas en la subsección 4.4.2,

en el sistema IR-n.

Todos estos experimentos han sido realizados sobre las dos colecciones descritas y utilizan tanto preguntas cortas como largas, excepto en casos en los que sólo tenga sentido uno de estos tipos.

5.3.2 Visualización de resultados

Como base para la presentación de los resultados de los experimentos se han empleado las medidas utilizadas en la conferencia CLEF (descritos en las subsecciones de la 5.2.7 a la 5.2.10). No obstante, se ha seleccionado la información más relevante de cada una de ellas y se han presentado de forma que facilite el proceso de comparación. Así para cada prueba se mostrarán los valores siguientes:

- Cobertura.
- Precisión a los 5, 10 20 30 y 200 documentos.
- Media de precisión interpolada (AvgP).

		Precisión a los N documentos					
	Cobertura	**5**	**10**	**20**	**30**	**200**	**AvgP**
Coseno	94.0	0.6000	0.5408	0.4582	0.4054	0.1826	0.4699

Tabla 5.8. Ejemplo de visualización de tabla de resultados

La tabla 5.8 muestra un ejemplo de visualización de los resultados. La primera columna indica una breve descripción del experimento, la segunda indica el valor de la cobertura, las 5 siguientes muestran la precisión a los 5, 10, 20, 30 y 200 documentos y la última muestra el valor de la media de precisión interpolada.

Además, para comparar los resultados entre las diferentes pruebas realizadas, se valorará principalmente la media de precisión interpolada (en adelante *AvgP*). Esto es debido a que es una medida que no sólo valora la precisión obtenida por el sistema, sino que también mide el hecho que los documentos relevantes sean recuperados en los primeros lugares. Esta es la medida utilizada

en las conferencias CLEF para comparar los resultados de los distintos sistemas participantes. Por ello, en algunos casos en los que se desee comparar los valores de precisión media de las distintas pruebas, se añadirá una columna adicional que mostrará valores de incremento o disminución de este último valor sobre otros, como se puede ver en la tabla 5.9. En este ejemplo, el segundo método tiene una precisión media superior en un 6,6% sobre el primero.

		Precisión a los N documentos						
	Cob	5	10	20	30	200	AvgP	% Δ
Cortas	94,0	0,6000	0,5408	0,4582	0,4054	0,1826	0,4699	0
Largas	95,6	0,6163	0,5612	0,4857	0,4367	0,1943	0,5009	+6,6

Tabla 5.9. Ejemplo de visualización de resultados con comparativa de medias de precisión

5.3.3 Experimento 1. Definición del sistema base.

Objetivo. El objetivo de este experimento es calcular los resultados que obtiene un sistema estándar de RI, el cual definimos como base. Los resultados obtenidos permitirán efectuar valoraciones de comparación con respecto a los resultados que obtenga el sistema IR-n en posteriores experimentos.

Descripción del experimento. Como sistema base se ha seleccionado el modelo del coseno. De las diferentes formulaciones de este modelo se ha seleccionado la indicada en (Kaszkiel et al., 1999), descrita en el capítulo 2.

Los experimentos se han realizado para las colecciones de LA-Times y EFE utilizando tanto preguntas cortas como largas.

Resultados obtenidos. Los resultados obtenidos se pueden ver en la tabla 5.10. Las dos primeras filas muestran los resultados obtenidos en la colección de LATimes y las dos últimas los obtenidos en la colección EFE, para preguntas cortas y largas respectivamente. En la última columna se muestra la variación porcentual

de la AvgP.

		Precisión a los N documentos						
	Cob	**5**	**10**	**20**	**30**	**200**	**AvgP**	**% Δ**
LATimes								
Cortas	94,2	0,3745	0,3255	0,2681	0,2255	0,0740	0,3723	0
Largas	95,8	0,4638	0,3766	0,2968	0,2475	0,0753	0,4711	+26,5
EFE								
Cortas	94,7	0,6082	0,5367	0,4561	0,4088	0,1859	0,4751	0
Largas	95,6	0,6163	0,5612	0,4837	0,4374	0,1942	0,5010	+5,4

Tabla 5.10. Resultados del modelo del coseno. Colecciones LATimes y EFE

Conclusiones. El primer elemento a destacar es el hecho de que los resultados obtenidos en cada una de las colecciones sea tan diferente a pesar de que se utilicen las mismas preguntas en ambos casos. Como se puede observar, cuando se utilizan las preguntas cortas en la colección LATimes se obtiene una precisión media del 37%, mientras en la de la agencia EFE la precisión obtenida es del 47%. Este hecho se debe a que la colección de la agencia EFE contiene una mayor cantidad de documentos relevantes que la LATimes. En la primera, el número total de documentos relevantes es de 2.694, mientras que en la de LATimes, es sensiblemente menor, conteniendo un total de 856 documentos relevantes.

El segundo elemento a destacar son los resultados obtenidos utilizando preguntas cortas y largas. La conclusión que se extrae es que los resultados son sensiblemente mejores cuando se utilizan las preguntas largas, ya que éstas aportan mayor cantidad de información sobre lo que se considera documento relevante.

Otro aspecto a considerar, es el motivo por el cual el beneficio de utilizar preguntas largas sobre las cortas es sensiblemente mayor en el caso de LATimes. Esto se debe a la dificultad que tiene el sistema de localizar en esta colección los relativamente pocos documentos relevantes utilizando poca información. Debido a esto, al aportar mayor información en la pregunta, se eleva considerablemente el número de documentos relevantes recuperados.

5.3.4 Experimento 2. Obtención del tamaño óptimo de los pasajes.

Objetivo. Determinar el tamaño de los pasajes, medido en número de frases, con el que se obtiene el mejor rendimiento del sistema.

Descripción del experimento. En este experimento se ha utilizado la medida $IR - n_{base}$ descrita en la subsección 4.2.2.

Estas pruebas se han realizado teniendo en cuenta los siguientes aspectos:

- El grado de solapamiento utilizado es de una frase.
- El tamaño de los pasajes ha sido un elemento que ha variado en cada una de las pruebas realizadas en este experimento.
- La medida de similitud utilizada es $IR - n_{base}$.

En primer lugar, se ha experimentado con incrementos de cinco frases, probando de cinco a cuarenta frases. Una vez realizadas estas pruebas, se seleccionó el intervalo con mejor resultado y se repitió el proceso en dicho intervalo con incrementos de una única frase. Finalmente se definió como tamaño óptimo aquel con el que se obtuvo mejor resultado.

Resultados obtenidos. En la tabla 5.11 pueden verse los resultados (en AvgP) obtenidos en ambas colecciones, utilizando diferentes tamaños de pasajes (5, 10, 15, 20, 25, 30 y 35). En cada fila se indica la colección y tipo de pregunta utilizada. Los resultados completos de este experimento se detallan en las tablas A.1, A.2, A.3 y A.4 incluidas en el anexo A.

Dado que alrededor de los pasajes de 10 frases se obtienen los mejores resultados, se repiten los experimentos para tamaños de 5 a 15 frases en incrementos de 1 frase en ambas colecciones y tipos de preguntas. La tabla 5.12 muestra un extracto de los resultados. Los resultados completos de este experimento pueden consultarse en las tablas A.5, A.6, A.7 y A.8 incluidas en el anexo A.

Conclusiones. Como se puede comprobar en la tabla 5.11, en tres de los cuatro modelos probados se obtienen mejores resultados

	Tamaño de pasaje en número de frases						
Preg.	**5**	**10**	**15**	**20**	**25**	**30**	**35**
LATimes							
Cortas	0,4622	**0,4896**	0,4667	0,4751	0,4652	0,4563	0,4493
Largas	0,4913	**0,5007**	0,4965	0,4993	0,4856	0,4841	0,4739
EFE							
Cortas	0,4969	**0,5052**	0,4963	0,4917	0,4908	0,4889	0,4885
Largas	**0,5085**	0,5007	0,4819	0,4728	0,4716	0,4709	0,4700

Tabla 5.11. AvgP para las colecciones LATimes y EFE utilizando diferentes tamaños de pasajes

	Tamaño de pasaje en número de frases						
Preg.	**5**	**6**	**7**	**8**	**9**	**10**	**11**
LATimes							
Cortas	0,4622	0,4780	0,4750	0,4744	0,4736	**0,4896**	0,4871
Largas	0,4913	**0,5160**	0,5012	0,4984	0,5040	0,5007	0,5035
EFE							
Cortas	0,4969	0,5081	0,5071	**0,5093**	0,5064	0,5052	0,5046
Largas	0,5085	0,5140	**0,5178**	0,5149	0,5075	0,5007	0,4967

Tabla 5.12. AvgP para las colecciones LATimes y EFE utilizando diferentes tamaños de pasajes

utilizando pasajes de diez frases. La evolución de los valores de AvgP en función del tamaño de pasaje puede verse en los gráficos 5.4 y 5.5, para las colecciones LATimes y EFE respectivamente. En estos gráficos se comprueba que una vez alcanzado el óptimo, el valor de AvgP decrece a medida que aumenta el tamaño de pasaje utilizado. Este decrecimiento depende del tipo de la colección. Como se puede ver en el caso de la agencia EFE, un vez alcanzado un tamaño de pasaje superior al tamaño que tienen la mayoría de documentos de la colección, apenas varía el valor del AvgP.

Los tamaños de pasajes para los que se obtienen mejores resultados depende del tipo de colección y del tipo de pregunta, aunque todos ellos se encuentran dentro de un intervalo reducido (pasajes de 6 a 10 frases).

En primer lugar, cabe destacar que el tamaño de pasaje óptimo es más pequeño en el caso de las preguntas largas (pasajes de 6 y 7 frases en las colecciones LATimes y EFE respectivamente) que para el caso de las preguntas cortas (pasajes de 10 y 8 frases para las colecciones LATimes y EFE respectivamente). Este hecho ya se apunta en Kaszkiel y Zobel (2001), donde en pruebas realizadas

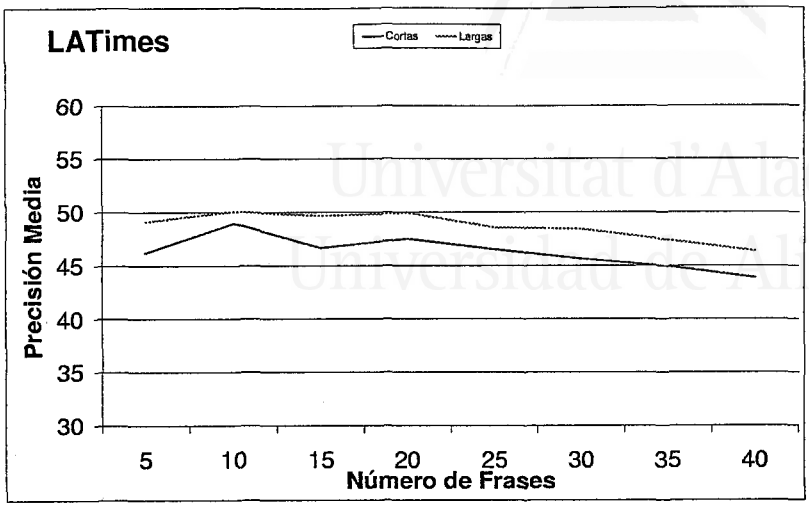


Figura 5.4. Valores de AvgP en función del tamaño del pasaje utilizado. Colección LATimes

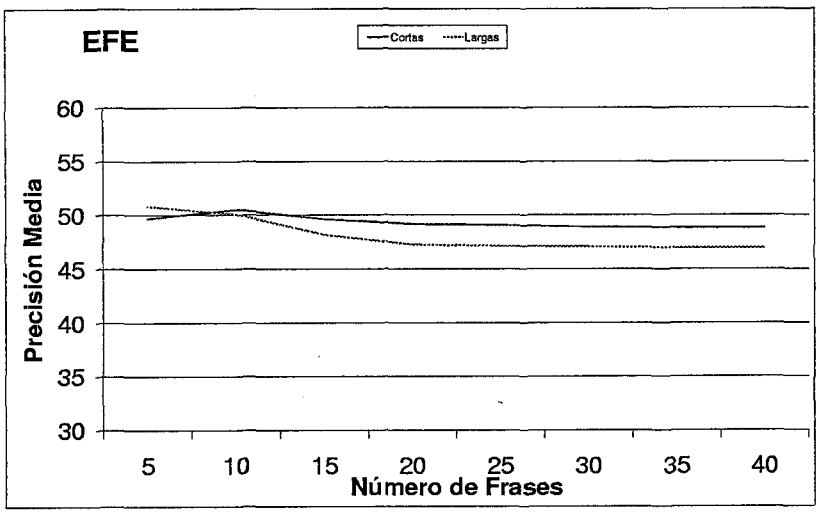


Figura 5.5. Valores de AvgP en función del tamaño del pasaje utilizado. Colección EFE

con el modelo de RP denominado *Arbitrary passages*, también se obtienen mejores resultados con pasajes de menor tamaño cuando se utilizan preguntas largas. Esto se debe a que al reducir el tamaño del pasaje, se facilita que aparezcan en el mismo concepto una de las ideas que se solicitan en las preguntas largas.

En segundo lugar cabe comparar los resultados obtenidos por el sistema IR-n frente a los obtenidos por el modelo del coseno. En la tabla 5.13 se muestra la comparación de los resultados obtenidos en ambas colecciones con preguntas cortas. Estos resultados indican que el modelo IR-n obtiene unos resultados sensiblemente superiores a los del modelo del coseno (un 31,5% en el caso de LATimes y un 7,2% en el caso de la agencia EFE).

		Precisión a los N documentos						
	Cob	**5**	**10**	**20**	**30**	**200**	**AvgP**	**% Δ**
LATimes								
Coseno	94,2	0,3745	0,3255	0,2681	0,2255	0,0740	0,3723	0
IR-n	95,2	0,5106	0,4170	0,3096	0,2624	0,0752	0,4896	+31,5
EFE								
Coseno	94,7	0,6082	0,5367	0,4561	0,4088	0,1859	0,4751	0
IR-n	95,1	0,6694	0,5898	0,4990	0,4463	0,1940	0,5093	+7,2

Tabla 5.13. Comparativa modelos coseno y sistema IR-n. Preguntas cortas

		Precisión a los N documentos						
	Cob	**5**	**10**	**20**	**30**	**200**	**AvgP**	**% Δ**
LATimes								
Coseno	95,8	0,4638	0,3766	0,2968	0,2475	0,0753	0,4711	0
IR-n	95,9	0,5234	0,4021	0,3117	0,2660	0,0768	0,5160	+9,5
EFE								
Coseno	95,6	0,6163	0,5612	0,4837	0,4374	0,1942	0,5010	0
IR-n	96,1	0,6612	0,5959	0,5071	0,4517	0,1967	0,5178	+3,4

Tabla 5.14. Comparativa modelos coseno y sistema IR-n. Preguntas largas

En la tabla 5.14 se muestra la comparación de los resultados obtenidos por el sistema IR-n y el modelo del coseno, en las colecciones de LATimes y agencia EFE cuando se utilizan preguntas largas. Con este tipo de pregunta el modelo IR-n también obtiene unos resultados sensiblemente superiores a los del modelo del

coseno en ambas colecciones (un 9,5% en la colección LATimes y un 3,4% en la colección EFE). Hay que destacar también, que el porcentaje de mejora es inferior en el caso de preguntas largas que cuando se utilizan las preguntas cortas. Esto se debe a que una pregunta larga amplía el rango de elementos a considerar dentro del documento, con lo que es difícil que se hallen contenidos todos en pasajes reducidos. No obstante, hay que indicar que en cualquiera de los casos, el resultado que obtiene el sistema siempre es superior a los obtenidos por el modelo del coseno.

Además, el motivo por el cual el incremento del AvgP es menor en el caso de las noticias de la agencia EFE, es que ésta es una colección de documentos más pequeños y de tamaño mucho más homogéneo. Por ello es más complicado que un sistema de RP aporte las ventajas propuestas en su modelo. Por otro lado en la colección LATimes se obtienen incrementos notables en el AvgP, ya que esta colección está formada por un conjunto de documentos de tamaño muy heterogéneo.

En los siguientes experimentos se utilizarán para cada una de las colecciones y tipos de preguntas los tamaños de pasajes que mejores resultados se han obtenido en este experimento, estos tamaños se muestran en la tabla 5.15.

	LATimes	EFE
Cortas	10	8
Largas	6	7

Tabla 5.15. Tamaños óptimos de pasajes por tipo de pregunta y colección

5.3.5 Experimento 3. Obtención del grado de solapamiento óptimo.

Objetivo. Por defecto, el sistema IR-n utiliza un grado de solapamiento 1. Esto implica que cada nuevo pasaje generado empieza en la segunda frase del pasaje anterior. Con este modelo se minimiza la posibilidad de que algunos documentos que contengan las palabras de la pregunta en frases consecutivas, puedan ser consi-

derados menos relevantes, al no hallarse dichas frases en el mismo pasaje.

Todas las pruebas realizadas hasta el momento han utilizado este grado de solapamiento. Evidentemente esto incrementa el coste temporal del proceso de búsqueda, ya que el número de pasajes diferentes a evaluar es mayor que si se utilizaran grados de solapamiento más elevados. En este experimento se plantea utilizar diferentes grados de solapamiento, que van desde el solapamiento mínimo de una frase, hasta el máximo que sería determinado por el tamaño del pasaje utilizado.

Descripción del experimento. En este experimento se plantea utilizar un mismo tamaño de pasaje para las dos colecciones y tipos de pregunta con la intención de poder comparar fácilmente los resultados obtenidos. El tamaño elegido es la media de los considerados óptimos en el experimento 2, que es el pasaje formado por 8 frases.

Se han seleccionado diversos grados de solapamiento desde una frase hasta 8 frases, éstos han sido 1, 2, 4 y 8. Es decir un grado de solapamiento 2 implica que el segundo pasaje empieza en la tercera frase del primer pasaje. Un grado de solapamiento 8 en este caso implicaría la no existencia del mismo, es decir, un pasaje empieza en la frase siguiente en la que termina el anterior pasaje.

Para poder evaluar el coste de tiempo que supone utilizar un grado de solapamiento menor, también se ha contabilizado el tiempo empleado por el sistema en la realización de cada una de las pruebas en las diferentes colecciones.

Resultados obtenidos. En la tablas 5.16 y 5.17 se muestran, para cada experimento, el valor de la media de precisión y el tiempo de sistema utilizado. Además, junto a cada uno de estos valores se muestra el porcentaje de decremento de los mismos con respecto al uso del grado de solapamiento 1. Los resultados completos de este experimento se detallan en las tablas A.9 y A.10 del anexo A.

Conclusiones. Al observar los resultados contenidos en las tablas 5.16 y 5.17 se puede comprobar que el uso de grados de

Sol.	AvgP	% Δ AvgP	Tiempo	% Δ Tiempo
LATimes				
1	0,4744	0	3:27.53	0
2	0,4649	-2.0	3:20.30	-3,4
4	0,4513	-4.9	3:19.85	-4
8	0,4477	-5.6	3:19.52	-4
EFE				
1	0,5093	0	2:08.97	0
2	0,5070	-0.5	2:07.56	-0,8
4	0,5057	-0.7	2:07.75	-0,8
8	0,5040	-1	2:07.61	-0,8

Tabla 5.16. Comparativa AvgP y Tiempo de sistema utilizando diferentes grados de solapamiento. Colecciones LATimes y EFE con preguntas cortas

Sol.	AvgP	% Δ AvgP	Tiempo	% Δ Tiempo
LATimes				
1	0.5025	0	14:29.76	0
2	0.4973	-1.0	14:00.54	-3.3
4	0.4548	-9.5	13:48.22	-4.7
8	0.4598	-8.5	13:42.62	-5.4
EFE				
1	0.5192	0	5:40.82	0
2	0.5151	-0.8	5:29.37	-3.2
4	0.5147	-0.9	5:27.30	-3.8
8	0.5135	-1.1	5:26.30	-4.1

Tabla 5.17. Comparativa AvgP y Tiempo de sistema utilizando diferentes grados de solapamiento. Colecciones LAtimes y EFE con preguntas largas

solapamiento mayores a uno disminuye el rendimiento del sistema (AvgP). Este decremento es mucho menor en el caso de la colección de LATimes. Esto se debe a que en la colección EFE los documentos son muy pequeños, con lo que el número de pasajes que no se procesan es mucho más pequeño que en el caso de la colección LATimes.

También hay que indicar que a nivel de coste temporal, este incremento del grado de solapamiento no supone apenas un descenso apreciable del tiempo de ejecución. Ello es debido a la forma en la que ha sido implementado el sistema IR-n. Así, este mayor grado de solapamiento incide únicamente en el número de pasajes a evaluar sobre una estructura ya cargada en memoria. Considerando todo el proceso de recuperación, la mayor parte del tiempo se utiliza en el acceso a disco para la obtención de información y su posterior carga en memoria, así como en la escritura

de los resultados. En consecuencia, teniendo en cuenta tanto el rendimiento del sistema como las escasas diferencias observadas en complejidad temporal, la mejor opción continúa siendo utilizar el grado de solapamiento 1.

5.3.6 Experimento 4. Aplicación de medidas de proximidad.

Objetivo. El objetivo de este experimento es determinar si se obtienen mejores resultados al utilizar el refinamiento del sistema $IR - n_{base}$, denominado $IR - n_{prox}$, que fue descrito en la subsección 4.4.1. La finalidad de este refinamiento es valorar el hecho que aparezcan las palabras consecutivas de la pregunta en la misma frase del documento. En la medida $IR - n_{prox}$ se fija el factor α, que determina el grado de incremento que se da al valor de similitud de un documento cuando sucede este hecho. En este experimento se pretende determinar para qué valores de α se obtienen mejores resultados.

Descripción del experimento. Se realizan pruebas utilizando los valores de α iguales a 1, 1'1, 1'2, 1'3 y 1'4. Se considera el factor 1 como un valor que define el modelo base, ya que en este caso no hay incremento de valoración en función de la aparición en la misma frase de palabras consecutivas.

Resultados obtenidos. En la tabla 5.18 se muestra la información resumida que contiene los valores de AvgP obtenidos en todas las pruebas, así como el incremento o decremento que se obtiene en cada una de ellas sobre el valor base ($\alpha = 1$). En el anexo A se incluyen las tablas A.11 y A.12 que muestran los resultados completos para ambas colecciones utilizando los factores indicados.

Conclusiones. La primera conclusión que se extrae de este experimento es que se obtienen ligeras mejorías en la media de precisión cuando los valores de α son pequeños. A medida que este valor aumenta, decrece el rendimiento observado. El elemento a destacar es que al utilizar valores de α iguales a 1,1 se obtiene una

	Valores de α							
Preg.	**1,1**	**% Δ**	**1,2**	**% Δ**	**1,3**	**% Δ**	**1,4**	**% Δ**
LATimes								
Cortas	0,4927	+0,6	0,4843	-1,1	0,4826	-1,4	0,4825	-1,5
Largas	0,5204	+0,3	0,5232	+0,8	0,5105	-1,5	0,5114	-1,4
EFE								
Cortas	0,5114	+0,4	0,5109	+0,3	0,5130	+0,7	0,5117	+0,5
Largas	0,5227	+0,4	0,5218	+0,2	0,5194	-0,2	0,5163	-0,8

Tabla 5.18. Aplicación medidas de proximidad en el sistema IR-n

mejora en todos los casos, aunque en algunos casos particulares se obtienen mejores resultados con valores superiores de α.

El estudio de las preguntas permite comprobar que, en general, este modelo obtiene mejores resultados cuando aparecen en ellas conceptos o entidades formadas por más de una palabra. Por ello se plantea la conveniencia de realizar un análisis y búsqueda de conceptos o entidades en la pregunta para poder aplicar este método con mejores resultados.

5.3.7 Experimento 5. Separación de las preguntas largas

Objetivo. Este experimento tiene como finalidad valorar los posibles beneficios que se obtienen al procesar una pregunta larga, mediante el proceso de una serie de subpreguntas cortas. Estas subpreguntas estarían formadas por cada una de las frases que forman la pregunta larga original. Así, este experimento se basa en la idea de dividir las preguntas largas en las frases que la forman y valorar la relevancia de cada documento en función de como respondan a una de ellas. Este modelo ha sido descrito en la subsección 4.4.3 de esta tesis.

Descripción del experimento. Para adecuar el modelo propuesto por el sistema IR-n a las preguntas del CLEF se realiza un tratamiento adicional. Esto se debe a que en muchos casos cada una de las frases que forman una pregunta larga, contienen referencias anáforicas a frases anteriores, por lo que pierden parte de su sentido cuando son tratadas de forma independiente. Dado que, en general, la parte de título y descripción de la pregunta

definen el contexto de la misma, se incorporan éstos a cada una de las preguntas lanzadas.

Así las tareas que se realizan para cada una de las preguntas son las siguientes:

- Se divide la parte narrativa de cada pregunta según las frases que la componen.
- Se generan tantas subpreguntas como frases componen la narrativa. Cada subpregunta contiene *título* + *descripción* + *frase de la narrativa*. El motivo de incorporar a todas las preguntas el título y la descripción se debe a que en muchos casos se requiere para centrar el contexto.
- Se procesan de forma separada cada subpregunta generada recuperando 5000 documentos.

Para obtener la similitud final del documento se han utilizado dos modelos diferentes. Éstos son los siguientes:

- Se otorga a cada documento, como puntuación, la suma de AvgP que ha obtenido en cada una de las subpreguntas.
- Se otorga a cada documento, como puntuación, la máxima AvgP que ha obtenido en cada una de las subpreguntas.

Un ejemplo de la descomposición de una pregunta sería el descrito en el (36):

(36) **Pregunta original**
Título El tratado de paz entre Israel y Jordania
Descripción Encontrar noticias que mencionen los nombres de los principales negociadores del tratado de paz en el Medio Oriente entre Israel y Jordania, y también documentos que den una información detallada sobre el tratado.
Narrativa
Frase 1 El 26 de octubre de 1996 se firmó un tratado de paz entre Israel y Jordania, abriendo nuevas posibilidades para las relaciones diplomáticas entre ambos países.
Frase 2 Los documentos relevantes proporcionarán detalles del tratado y/o mencionarán a los participantes

principales de las negociaciones.

Subpreguntas

Pregunta 1 El tratado de paz entre Israel y Jordania. Encontrar noticias que mencionen los nombres de los principales negociadores del tratado de paz en el Medio Oriente entre Israel y Jordania, y también documentos que den una información detallada sobre el tratado. El 26 de octubre de 1996 se firmó un tratado de paz entre Israel y Jordania, abriendo nuevas posibilidades para las relaciones diplomáticas entre ambos países.

Pregunta 2 El tratado de paz entre Israel y Jordania. Encontrar noticias que mencionen los nombres de los principales negociadores del tratado de paz en el Medio Oriente entre Israel y Jordania, y también documentos que den una información detallada sobre el tratado. Los documentos relevantes proporcionarán detalles del tratado y/o mencionarán a los participantes principales de las negociaciones.

Estas pruebas se han realizado teniendo en cuenta los siguientes aspectos:

- El grado de solapamiento utilizado es de una frase.
- El tamaño de los pasajes utilizado ha sido el óptimo para cada colección al utilizar preguntas cortas.
- La medida de similitud utilizada es $IR - n_{prox}$ (con el valor $\alpha = 1,1$).

Resultados. En la tabla 5.19 se describen los resultados obtenidos para cada una de las colecciones con los dos modelos de tratamiento de las preguntas separadas. Ambos se comparan con el lanzamiento de preguntas largas de forma completa. Se definen de la siguiente forma:

- *Completa(Com.)*: Es el sistema IR-n utilizando la pregunta larga sin dividir.

- *Suma*: Se define la similitud del documento como la suma de las similitudes de dicho documento para cada una de las preguntas en las que se separa la pregunta original.
- *Máximo(Max.)*: Se define la similitud del documento como la máxima que ha obtenido en cada una de las preguntas separadas.

		Precisión a los N documentos						
	Cob	**5**	**10**	**20**	**30**	**200**	**AvgP**	**% Δ**
LATimes								
Com.	95,9	0,5362	0,4128	0,3223	0,2674	0,0773	**0,5204**	0
Suma	95,8	0,5234	0,4234	0,3234	0,2695	0,0783	0,5168	-0,7
Max.	95,9	0,5149	0,4298	0,3266	0,2596	0,0760	0,5082	-3,2
EFE								
Com.	96,4	0,6694	0,5918	0,5061	0,4558	0,1981	0,5227	0
Suma	96,1	0,6735	0,6014	0,5122	0,4592	0,1995	**0,5284**	+1,1
Max.	95,6	0,6776	0,5898	0,5102	0,4558	0,1945	0,5123	-2

Tabla 5.19. Comparativa modelos de separación de preguntas largas. Colecciones LATimes y EFE

Conclusiones. Los primeros datos a destacar son fundamentalmente dos. En primer lugar, el modelo *Suma* obtiene sensibles mejoras de resultados sobre el modelo *Máximo*. Esto se debe fundamentalmente a que al dividir las preguntas en subpreguntas (cada una formada por una frase), puede ocurrir que algunas de ellas estén formadas por un mayor número de palabras o por palabras más discriminantes que las contenidas en otras subpreguntas. Este hecho provoca que las puntuaciones obtenidas por los documentos para una frase en general sean sensiblemente mayores que los obtenidos en otra. Esto puede producir que documentos algo relevantes a una subpregunta, sean considerados más relevantes que documentos muy relevantes a otra subpregunta. Esto no ocurre de la misma forma al utilizar la suma de la relevancia de cada subpregunta.

En segundo lugar cabe destacar que el modelo de *Suma* obtiene mejores resultados que el modelo basado en procesar la pregunta completa (modelo base) en la colección EFE. No ocurre así en la colección LATimes. Esto se debe a que en las preguntas utilizadas en la colección LATimes se obtienen resultados sensiblemente

mejores cuando se utiliza la parte narrativa de la pregunta que la parte de título y descripción. No ocurre de la misma forma en la colección EFE. Dado que en cada subpregunta se incorpora el título y descripción, los resultados van a ser dependientes de los resultados que se obtienen utilizando solamente el título y la descripción de la pregunta.

Como conclusiones adicionales cabe resaltar que este modelo permite en algunos casos mejorar los resultados de un proceso de la pregunta completa dentro de un sistema de RP. Otro aspecto a destacar y que formará parte de trabajos que se deben realizar en un futuro, es estudiar la posibilidad de realizar un análisis más completo de las preguntas largas, de forma que se extraiga cada una de las ideas fundamentales que deben tener los documentos relevantes. Este análisis debería incorporar la resolución de posibles anáforas y correferencias dentro de la pregunta, de forma que cada frase tuviera sentido en sí misma.

También se debe utilizar algún mecanismo que permita normalizar los resultados obtenidos por cada una de las subpreguntas, de forma que tengan el mismo peso cuando se determina la relevancia de los documentos.

5.3.8 Experimento 6. Expansión de la pregunta.

Objetivo. Comprobar las ventajas que pueden aportar las técnicas de expansión de la pregunta al modelo del sistema IR-n, de acuerdo a la descripción realizada en la subsección 4.4.2 de esta tesis, y que se basa en las técnicas de análisis local.

Descripción del experimento. Se ha utilizado el modelo $IR-n_{prox}$ (con el valor $\alpha = 1,1$) y grado de solapamiento igual a uno, incorporando las técnicas de expansión de la pregunta tanto para preguntas cortas como largas.

Resultados. En las tablas 5.20 y 5.21 se muestran los resultados obtenidos por el modelo IR-n sin expansión de la pregunta (*Sin*) y con expansión (*Con*).

		Precisión a los N documentos						
	Cob.	5	10	20	30	200	AvgP	% Δ
LATimes								
Sin	95,1	0,5319	0,4170	0,3117	0,2603	0,0751	0,4927	0
Con	97,7	0,5064	0,4319	0,3245	0,2730	0,0799	0,5035	+2,2
EFE								
Sin	95,3	0,6653	0,6000	0,5010	0,4503	0,1952	0,5114	0
Con	94,4	0,6490	0,5898	0,5092	0,4524	0,1918	0,5300	+3,6

Tabla 5.20. Comparativa modelos sin y con expansión Preguntas cortas

		Precisión a los N documentos						
	Cob.	5	10	20	30	200	AvgP	% Δ
LATimes								
Sin	95,9	0,5362	0,4128	0,3223	0,2674	0,0773	0,5204	0
Con	96,5	0,5234	0,4404	0,3202	0,2631	0,0767	0,5111	-1,77
EFE								
Sin	96,4	0,6694	0,5918	0,5061	0,4558	0,1981	0,5227	0
Con	95,8	0,6735	0,6122	0,5214	0,4653	0,1961	0,5432	+3,9

Tabla 5.21. Comparativa modelos sin y con expansión. Preguntas largas

Conclusiones. Al observar los resultados cabe indicar que la mejora obtenida es sensible en la colección EFE, (sobre un 4%), mientras que en la colección LATimes es pequeña en el caso de preguntas cortas, e incluso empeora los resultados en caso de preguntas largas. Este hecho se debe a que en la colección EFE al utilizar la pregunta original, se obtienen mejores resultados de precisión a los cinco documentos recuperados, que son a partir de los cuales se obtienen los términos a añadir a la pregunta original.

Al valorar estos resultados se concluye que es conveniente utilizar técnicas de expansión de la pregunta, ya que la mejora puede ser significativa.

5.3.9 Análisis de los resultados de entrenamiento

La experimentación ha permitido definir de qué forma se debe aplicar el sistema IR-n para conseguir los mejores resultados.

Sobre los experimentos realizados con el modelo $IR - n_{base}$ cabe destacar:

1. Se han definido los tamaños de pasajes a utilizar con cada colección y tipo de pregunta con el que se obtienen mejores resultados.

2. Se ha comprobado que se consiguen mejores resultados cuando el grado de solapamiento es el mínimo (1 frase). Además, se ha demostrado que gracias a la forma en la que se ha implementado el sistema IR-n, el utilizar este grado de solapamiento incrementa muy poco el tiempo total del proceso de búsqueda (sobre un 4% sobre el modelo que no utiliza ningún solapamiento de pasajes).

Por lo que se refiere a los refinamientos definidos sobre el sistema IR-n cabe indicar:

1. El modelo $IR-n_{prox}$ (con $\alpha = 1,1$), el cual valora la proximidad de aparición de las palabras de la pregunta dentro de cada pasaje, permite mejorar los resultados del sistema $IR-n_{base}$. Aunque esta mejora no es significativa (sobre un 0,4%), sí que se produce en todos los tipos de preguntas y colecciones usadas en el proceso de entrenamiento. Por ello, y dado que el uso de esta medida no supone un incremento sensible de complejidad, el modelo $IR-n_{prox}$ será el modelo utilizado.
2. El tratamiento de forma separada de las preguntas largas, como si fuesen varias preguntas cortas, también ha permitido una mejora de los resultados obtenidos en las pruebas realizadas sobre las preguntas largas. No obstante, esta mejora sólo se produce en una de las colecciones (EFE). Dado que la evaluación final se realizará sobre esta colección, este modelo (*Suma*) será uno de los utilizados.
3. El incorporar técnicas de expansión de la pregunta permite mejorar de forma muy significativa (hasta un 4%) los resultados del sistema IR-n.

5.4 Evaluación del sistema IR-n. Conferencia CLEF-2002

La evaluación final del sistema IR-n presentado en este trabajo se llevó a cabo mediante su participación en la tarea monolingüe español del CLEF-2002. No se han podido realizar pruebas sobre

el idioma inglés en el ámbito de esta conferencia, ya que éste sólo se podía utilizar en la tarea de RI bilingüe y multilingüe.

Las principales ventajas de esta evaluación son las siguientes:

- Para realizar una evaluación independiente del entrenamiento del sistema, se necesitaba una colección de test de alta calidad que fuese diferente a la ya utilizada en este proceso de entrenamiento.
- Esta participación permite que el sistema sea evaluado por personas independientes, ajenas a las que desarrollaron el trabajo y con experiencia en la evaluación de sistemas de RI.
- Al utilizar diferentes sistemas de RI sobre las mismas colecciones de test, siendo todos ellos evaluados en base a los mismos criterios de corrección, se puede comparar el rendimiento del sistema IR-n con los sistemas de RI más importantes.

En los siguientes apartados se detallan las especificaciones de la tarea monolingüe CLEF-2002, se analizan los resultados obtenidos y se compara el rendimiento del sistema con todos los demás participantes en esta tarea.

5.4.1 Descripción de la tarea

En la tarea monolingüe español de la edición CLEF-2002 se utilizó la misma colección EFE descrita en el proceso de entrenamiento. La colección de preguntas está formada por 50 nuevas preguntas numeradas, de la 91 a la 140. A diferencia de la colección de preguntas utilizadas en la edición del año anterior, todas tenían al menos un documento que era relevante a las mismas.

Se permitió que cada participante realizara como máximo cuatro pruebas. La única limitación era que al menos una de estas pruebas hiciera uso únicamente del título y descripción, para permitir la comparación de todos los sistemas.

5.4.2 Descripción de las pruebas oficiales realizadas.

Cuatro fueron las pruebas que se enviaron a la organización del CLEF. Dos de ellas emplearon el título y descripción de la pregunta y las otras dos además incorporaron la parte narrativa de

la misma. En estas pruebas se utilizaron los modelos que mejores resultados se habían obtenido en el proceso de experimentación.

Las características comunes a todas las pruebas realizadas son las siguientes:

1. Como medida de similitud se ha utilizado la $IR - n_{prox}$, con un factor α de 1,1.
2. El grado de solapamiento de los pasajes es de uno.

Las características específicas de cada prueba, así como la definición de un nombre que las identifica se describen a continuación.

Pruebas utilizando sólo título y descripción de la pregunta.

Se realizaron dos pruebas: *IR-n P1* y la *IR-n P2*. En ambas se utilizaron pasajes formados por 8 frases, dado que con éste se obtuvieron los mejores resultados en la fase de experimentación. La diferencia entre ambas reside en el hecho de que en la prueba *IR-n P2* se utilizaron técnicas de expansión de la pregunta.

Pruebas utilizando la pregunta completa.

Se realizaron dos pruebas: *IR-n P3* y la *IR-n P4*. En la primera de ellas se utilizaron pasajes formados por 7 frases, tamaño de pasaje con el que se obtuvieron los mejores resultados en la fase de experimentación con este tipo de preguntas. En la segunda se utilizo el modelo de preguntas separadas con pasajes de 8 frases de tamaño.

5.4.3 Resultados obtenidos por el Sistema IR-n.

Los resultados oficiales de estas cuatro pruebas se muestran en la tabla 5.22 para las pruebas realizadas en preguntas cortas y en la 5.23 para las preguntas largas.

Estos resultados confirman los resultados obtenidos en el proceso de entrenamiento realizado. Es decir, se obtienen mejores resultados al utilizar el modelo de expansión de la pregunta y al utilizar las preguntas largas de forma separada. En ambos casos los porcentajes de mejora son similares a los obtenidos en la fase de entrenamiento.

		Precisión a los N documentos						
	Cob	5	10	20	30	200	AvgP	% Δ
IR-nP1	90,1	0,6800	0,5820	0,5140	0,4620	0,1837	0,4684	0
IR-nP2	91,8	0,6920	0,5920	0,5190	0,4667	0,2018	0,4980	+6,3

Tabla 5.22. Resultados oficiales CLEF-2002. Tarea monolingüe español. Pruebas sistema IR-n preguntas cortas

		Precisión a los N documentos						
	Cob	5	10	20	30	200	AvgP	% Δ
IR-nP3	91,8	0,7120	0,6120	0,5380	0,4867	0,1936	0,4976	0
IR-nP4	92,6	0,7200	0,6380	0,5600	0,4813	0,1898	0,5067	+1,8

Tabla 5.23. Resultados oficiales CLEF-2002. Tarea monolingüe español. Pruebas sistema IR-n preguntas largas

5.4.4 Comparación con otros sistemas

Una vez comparadas las cuatro pruebas realizadas, cabe estudiar los resultados obtenidos por el sistema IR-n y el resto de sistemas participantes. En la tabla 5.24 se muestran los mejores resultados obtenidos por cada sistema que se presentaron a la tarea monolingüe español utilizando las preguntas formadas por el título y la descripción. En la tabla 5.25 se muestran los resultados para las pruebas que incluían la parte narrativa de las preguntas. Ambas tablas ordenan los sistemas en función de la media de precisión interpolada (AvgP).

5.4.5 Análisis de los resultados obtenidos por el sistema IR-n en el CLEF-2002

Al estudiar las tablas 5.24 y 5.25 se puede comprobar que el sistema IR-n ha quedado clasificado, según el AvgP, en buen lugar tanto al utilizar preguntas cortas (quinto de trece participantes) como largas (segundo de cuatro participantes).

Otro aspecto a destacar es la comparación del sistema IR-n con la media de todos los sistemas participantes. Esta comparativa puede verse en la tabla 5.26. En esta tabla se compara la media de precisión obtenida por todos los sistemas, con los mejores resultados en cada uno de los casos por el sistema IR-n.

Los resultados del sistema IR-n mejoran en un 12,6% los resultados de la media de todos los sistemas participantes en las

		Precisión a los N documentos					
	Cob	**5**	**10**	**20**	**30**	**200**	**AvgP**
U.Neuchatel (Savoy, 2002)	93,1	0,6920	0,6200	0,5350	0,4787	0,2056	0,5441
U.Berkeley (Chen, 2002)	93,7	0,6600	0,6020	0,5200	0,4780	0,2096	0,5338
U. Johns Hopkins (McNamee y Mayfield, 2002)	93,2	0,6120	0,5920	0,5090	0,4700	0,2056	0,5192
Thomson L&R (Moulinier y Molina-Salgado, 2002)	89,7	0,6600	0,5960	0,5130	0,4593	0,1966	0,4993
IR-n (Llopis et al., 2002g)	91,8	0,6920	0,5920	0,5190	0,4667	0,2018	0,4980
Hummingbird (Thomlinson, 2002)	89,7	0,6760	0,5900	0,5150	0,4687	0,1988	0,4909
U.Exeter (Lam-Alesina y Jones, 2002)	88,2	0,6320	0,5680	0,4960	0,4387	0,1872	0,4745
U.Amsterdam (Monz et al., 2002)	86,2	0,6440	0,5480	0,4720	0,4207	0,1812	0,4734
Océ (Brand y Brünner, 2002)	88,6	0,6520	0,5760	0,5050	0,4367	0,1905	0,4557
U.La Coruña (Vilares et al., 2002)	87,4	0,4600	0,4540	0,4050	0,3693	0,1645	0,3608
U.Tolouse Sin ref,	72,2	0,4360	0,4100	0,3730	0,3380	0,1424	0,3305
U.Salamanca (Zazo et al., 2002)	81,5	0,4400	0,3940	0,3460	0,3100	0,1494	0,3143
City University (McFarlane, 2002)	68,4	0,3760	0,3540	0,3130	0,2820	0,1280	0,2552

Tabla 5.24. Resultados oficiales CLEF-2002. Tarea monolingüe español. Colección EFE con preguntas cortas. Sistemas ordenados por precisión media

		Precisión a los N documentos					
	Cob	**5**	**10**	**20**	**30**	**200**	**AvgP**
U. de Neuchatel	93,4	0,7760	0,6920	0,5960	0,5307	0,2128	0,6051
IR-n	92,6	0,7200	0,6380	0,5600	0,4813	0,1898	0,5067
U. La Coruña	92,3	0,5480	0,5280	0,4760	0,4327	0,1831	0,4448
U. Salamanca	89,1	0,5360	0,4900	0,4310	0,3967	0,1697	0,4051

Tabla 5.25. Resultados oficiales CLEF-2002. Tarea monolingüe español. Colección EFE con preguntas largas.

	AvgP	% Δ
Preguntas Cortas		
Media sistemas participantes	0.4422	0
IR-n P2	0.4980	+12.6
Preguntas Largas		
Media sistemas participantes	0.4904	0
IR-n P4	0.5067	+3,3

Tabla 5.26. Comparativa sistema IR-n, sin y con expansión de la pregunta. Colección EFE

preguntas cortas y en un 3,3% en el caso de las preguntas largas. En este último caso la diferencia sobre la media es sensiblemente menor, ya que el modelo utilizado en la conferencia CLEF por el sistema IR-n no incorporaba las técnicas de expansión de la pregunta.

Otro aspecto a destacar es el fenomenal comportamiento del sistema IR-n en la tarea de recuperar rápidamente los documentos relevantes. Si valoramos la precisión a los cinco documentos recuperados, se puede comprobar que es el sistema IR-n el que obtiene los mejores resultados. Si ordenamos por este aspecto la clasificación de los sistemas sería la descrita en la tabla 5.27.

Sistema	Precisión a los 5 documentos
IR-n	0.6920
U. de Neuchatel	0.6920
Hummingbird	0.6760
Thomson L&R	0.6600
U. de Berkeley	0.6600
Océ	0.6520
U. Amsterdam	0.6440
U. Exeter	0.6320
U. Johns Hopkins	0.6120
U. La Coruña	0.4600
U. Salamanca	0.4400
U. Tolouse	0.4360
City University	0.3760

Tabla 5.27. Resultados oficiales CLEF-2002. Tarea monolingüe español. Colección EFE con preguntas cortas. Sistemas ordenados por precisión a los 5 documentos relevantes

5.5 Comparativa del sistema IR-n con otros sistemas de Recuperación por Pasajes

Una vez se ha evaluado de forma externa el sistema IR-n frente a otros sistemas de RI mediante la participación en la conferencia CLEF-2002, se plantea una nueva evaluación del mismo frente a los sistemas de RP más conocidos con el objetivo de comparar y contrastar sus resultados. Las características de todos los modelos de RP evaluados en esta sección han sido previamente descritas en el capítulo 3 de esta tesis.

La evaluación realizada se compone de dos pruebas:

1. Se han implementado algunos modelos de RP y se han evaluado junto al sistema IR-n utilizando la colección LATimes y la colección de preguntas del CLEF-2002.
2. Se ha realizado una prueba con el sistema IR-n con la colección de test utilizada en Kaszkiel y Zobel (2001) y se han comparado los resultados obtenidos frente a los sistemas de RP evaluados en este trabajo. En esta comparativa se utiliza la colección *Federal Register* empleada en la conferencia TREC-4.

En las siguientes subsecciones se describirán las colecciones de test y ,los resultados obtenidos en cada una de las pruebas realizadas para posteriormente extraer las conclusiones obtenidas.

5.5.1 Experimento colección LATimes del CLEF-2002

La colección de test de este experimento se compone de:

- **Colección de documentos.** La colección de documentos es la LATimes (utilizada en la fase de experimentación descrita en este mismo capítulo).
- **Colección de preguntas.** Se ha utilizado la colección de preguntas de la CLEF-2002, tanto en su versión corta como larga. Esta colección es diferente a la que se utilizó en el entrenamiento del sistema IR-n (colección de preguntas de la CLEF-2001). La colección de preguntas está formada por 42 preguntas que al menos tienen un documento relevante en la colección LATimes.

Para realizar esta comparativa se han implementado los siguientes modelos de segmentación del documento en pasajes:

1. *Párrafos* (Salton et al., 1993). Cada pasaje se forma con cada uno de los párrafos del documento.
2. *Pages 1Kb* (Moffat et al., 1993). Se forman los pasajes con la unión de párrafos consecutivos hasta formar pasajes de al menos 1Kb.
3. *Pages 2Kb* (Zobel et al., 1995). Se forman los pasajes con la unión de párrafos consecutivos hasta formar pasajes de al menos 2Kb.
4. *Bounded Paragraphs* (Callan, 1994). Se forman los pasajes con la unión de párrafos consecutivos hasta formar pasajes formados por entre 50 y 200 palabras.
5. *TextTiling* (Hearst, 1994). Se forman los pasajes con la unión de párrafos consecutivos utilizando el algoritmo TextTiling.

Para cada uno de ellos se ha utilizado tanto, en el proceso de indexación como en el de búsqueda, el sistema de RI SMART, basado en el modelo del coseno. La definición del modelo del sistema IR-n utilizada en esta prueba ha sido la que mejores resultados obtuvo en la fase de experimentación con la colección LATimes, cuyas principales características son las siguientes:

- La medida utilizada es la $IR - n_{prox}$ con α igual a 1,1.
- El grado de solapamiento utilizado es de una frase.
- El tamaño de los pasajes utilizado ha sido de 10 frases para las preguntas cortas y 6 para las largas.

Las tablas 5.28 y 5.29 se muestran los resultados obtenidos para preguntas cortas y largas respectivamente. Además de los resultados de los sistemas de RP indicados, también se muestran los resultados obtenidos por la aplicación del modelo del coseno sobre el documento completo. En ambas tablas los sistemas se hallan ordenados por el valor de la media de precisión.

Como se puede ver en ambas tablas, el sistema IR-n obtiene mejores resultados que el resto de sistemas evaluados tanto utilizando preguntas cortas como largas. Cabe destacar que el modelo

basado en el documento completo obtiene buenos resultados cuando se utilizan las preguntas largas, pero pobres cuando se utilizan las cortas. Por último, se observa que en esta colección al estar formado sus documentos por párrafos muy pequeños, el modelo que se basa en los mismos para definir los pasajes obtiene unos resultados muy pobres, sólo mejores que el modelo que aplica el algoritmo *TextTiling*. No obstante, el modelo que une párrafos formando pasajes de aproximadamente 1Mb, mejora considerablemente estas propuestas al generar pasajes de tamaño mayor y homogéneo.

		Precisión a los N documentos					
	Cob.	**5**	**10**	**20**	**30**	**200**	**AvgP**
IR-n	86,5	0,4190	03405	0,2595	0,2159	0,642	**0,3529**
Pages 1Mb	89,5	0,3190	0,2929	0,2476	0,2206	0,0696	0,3160
Boundary Paragraphs	90,0	0,3476	0,3095	0,2560	0,2198	0,0708	0,3043
Pages 2Mb	91,0	0,3048	0,2881	0,2440	0,2135	0,0710	0,2792
Documento completo	77,9	0,3190	0,2595	0,1917	0,1579	0,0521	0,2155
Párrafos	77,1	0,2667	0,2452	0,1857	0,1643	0,0570	0,2083
TextTiling	81,2	0,2619	0,1976	0,1583	0,1286	0,0536	0,1551

Tabla 5.28. Comparativa de modelos de RP. Colección LATimes preguntas cortas

		Precisión a los N documentos					
	Cob.	**5**	**10**	**20**	**30**	**200**	**AvgP**
IR-n	96,5	0,5571	0,4381	0,3452	0,2841	0,0814	0,4631
Documento completo	94,5	0,4571	0,3643	0,2845	0,2444	0,0774	0,3826
Pages 1Mb	92,7	0,4190	0,3381	0,2810	0,2532	0,0743	0,3726
Boundary Paragraphs	93,1	0,4286	0,3571	0,2821	0,2548	00,0762	0,3559
Pages 2Mb	93,9	0,3524	0,3333	0,2810	0,2389	0,0768	0,3398
Párrafos	81,4	0,2952	0,2667	0,2071	0,1778	0,0624	0,2466
TextTiling	82,3	0,2619	0,2167	0,1619	0,1397	0,0531	0,1569

Tabla 5.29. Comparativa de modelos de RP. Colección LATimes preguntas largas

5.5.2 Experimento colección Federal Register del TREC-4

La colección de test de este experimento se compone de:

- **Colección de documentos.** La colección de test utilizada para este experimento ha sido la de los documentos del Federal Register que se hallan en los discos 1 y 2 de la colección del TREC. Esta colección fue utilizada en la edición del TREC-4. Esta colección está formada por documentos de tamaño muy heterogéneo (desde 93 bytes hasta 2,5 Mb). Las principales características de esta colección se muestran en la tabla 5.30.
- **Colección de preguntas.** Las preguntas que se han utilizado en este experimento es un subconjunto de las empleadas en la edición del TREC-4 (21 preguntas de la 51-100). Este conjunto de preguntas está formada por aquéllas para las que existe al menos un documento relevante en la colección. En estos experimentos se han utilizado también las versiones corta y larga de las preguntas.

Disco	Colección de documentos	Tamaño (MB)	Num. Docs.	Media Term. /Doc.
1	Federal Register (1988)	260	25.960	1315,9
2	Federal Register(1988)	209	19.860	1378,1

Tabla 5.30. Colecciones de documentos Federal Register

En esta prueba se toman como base los resultados descritos en (Kaszkiel y Zobel, 2001). En este artículo se comparan diversos modelos de RID como son los modelos del *coseno* y del *coseno pivotado*, y distintos sistemas de RP entre los que se encuentran algunos de los utilizados en el experimento anterior (*Párrafos*, *TextTiling* y *Pages 2Kb*), así como dos modelos de ventana adicionales (*Sliding windows* y *Arbitrary passages*). Las principales características de estos dos modelos adicionales son:

1. *Sliding Windows* (Callan, 1994). Cada pasaje se forma con un número determinado de palabras consecutivas (en este caso 350). Cada pasaje se solapa con el anterior cada 175 palabras.

2. *Arbitrary Passages* (Kaszkiel y Zobel, 2001). Se forman pasajes solapados de diferentes tamaños (25, 50 y 100 palabras consecutivas). Posteriormente se asigna a cada documento la similitud del mejor pasaje.

Dado que en el artículo no se describen las *listas de parada* utilizadas, ni la parte de los términos que se utilizan en la indexación, y con el objeto de comprobar si los resultados que se obtienen son comparables, se ha realizado una prueba previa en la que se utiliza el modelo del coseno utilizando las mismas condiciones de indexación que posteriormente se utilizarán con el sistema IR-n. Posteriormente, se comparan los resultados obtenidos en esta prueba con los descritos en el artículo utilizando el mismo modelo. Estos resultados se muestran en la tabla 5.31. En la misma se puede comprobar que los resultados obtenidos de forma experimental y los descritos en el artículo son muy similares (sobre un 1,4% de diferencia en el AvgP y valores similares de precisión).

	Precisión a los N documentos						
	5	**10**	**20**	**30**	**200**	**AvgP**	**% Δ**
coseno artículo	0,0857	0,0857	0,0667	0,0603	0,0267	0,1283	0
coseno experimental	0,0857	0,0810	0,0738	0,0698	0,0310	0,1265	-1,4

Tabla 5.31. Comparativa de modelos del coseno, obtenidos de forma experimental y según (Kaszkiel y Zobel, 2001)

En esta prueba se ha empleado el sistema IR-n con la misma configuración descrita en el experimento anterior.

La tabla 5.32 muestra los resultados obtenidos por todos los sistemas utilizando preguntas cortas. En la tabla no se hace constar el valor de la cobertura obtenida por cada sistema ya que esta información no está disponible en el artículo. En ella los sistemas aparecen ordenados en función de la precisión media obtenida.

Por otra parte, en la tabla 5.33 pueden verse los resultados obtenidos por todos los sistemas utilizando preguntas largas. En esta tabla los sistemas también aparecen ordenados en función de la precisión media obtenida.

	Precisión a los N documentos					
	5	**10**	**20**	**30**	**200**	**AvgP**
Arbitrary-passages	0,1905	0,1619	0,1286	0,1175	0,0424	0,2960
Sliding windows	0,1714	0,1476	0,1214	0,1063	0,0383	0,2701
IR-n	0,1905	0,1286	0,1024	0,0905	0,03690	0,2495
Párrafos	0,1333	0,1143	0,0881	0,0762	0,0310	0,2327
Pages	0,1810	0,1429	0,1143	0,0905	0,0355	0,2250
Documento completo coseno pivotado	0,1619	0,1429	0,1024	0,0937	0,0329	0,2075
TextTiling	0,1238	0,1238	0,1048	0,0968	0,0,0367	0,1985
Documento completo coseno	0,0857	0,0857	0,0667	0,0603	0,0267	0,1283

Tabla 5.32. Comparativa de sistemas de RI. Colección Federal Register. Sistemas ordenados por precisión media

	Precisión a los N documentos					
	5	10	20	30	200	AvgP
Arbitrary-passages	0,3238	0,2667	0,2119	0,1873	0,0724	0,3405
Sliding windows	0,2952	0,2571	0,2214	0,1841	0,0705	0,3245
Pages	0,2857	0,2524	0,2119	0,1810	0,0681	0,3143
IR-n	0,2800	0,2133	0,1533	0,1356	0,0513	0,3003
Documento completo coseno	0,2286	0,2238	0,1762	0,1508	0,0712	0,2928
Párrafos	0,2857	0,2190	0,1905	0,1683	0,0645	0,2790
Documento completo coseno pivotado	0,1810	0,1571	0,1357	0,1365	0,00533	0,2417
TextTiling	0,2667	0,2286	0,1952	0,1714	0,0674	0,2358

Tabla 5.33. Comparativa Federal Register preguntas largas, ordenados por precisión media

Al observar los resultados de ambas tablas, se comprueba que son similares a los de la prueba realizada sobre la colección LA-Times. El modelo RID del coseno obtiene buenos resultados al utilizar preguntas largas y malos con preguntas cortas. El sistema *TextTiling* es en general el que peores resultados obtiene. Los modelos de ventana con solapamiento que se incorporan en esta comparativa obtienen buenos resultados. Ambos sistemas obtienen mejores resultados, sobre todo el *Arbitrary passages*, aunque también hay que destacar que es el más complejo de los tres.

Aunque el sistema IR-n no mejora a estos dos a nivel de precisión media, es el mejor sistema a nivel de precisión a los cinco documentos recuperados, como se puede observar en la tabla 5.34. Esto indica que el sistema IR-n obtiene la mejor precisión cuan-

do se han recuperado pocos documentos, de la misma forma que ocurría en la evaluación realizada en el CLEF-2002.

Sistema	Precisión a los 5 documentos
IR-n	0,1905
arbitrary-passages	0,1905
pages	0,1810
sliding windows	0,1714
coseno pivotado	0,1619
párrafos	0,1333
texttiling	0,1238
coseno	0,0857

Tabla 5.34. Comparativa de sistemas de RI. Colección Federal Register. Sistemas ordenados por la precisión a los cinco documentos recuperados

5.5.3 Conclusiones de la comparativa

Se pueden extraer varias conclusiones de los resultados obtenidos en las pruebas realizadas.

- La primera conclusión es que los sistemas de RP funcionan mejor que los sistemas RID, sobre todo cuando se utilizan preguntas cortas. Esto se debe a que los sistemas de RP favorecen los documentos que tienen en pasajes reducidos varios términos muy relacionados que suelen formar parte de las preguntas cortas. Los sistemas RID mejoran sus resultados cuando las preguntas son de mayor tamaño.
- Los sistemas de RP de ventana que utilizan solapamiento de pasajes (entre los que se encuentra el sistema IR-n) son los que mejores resultados obtienen en las dos pruebas. Esto se debe a que estos modelos evitan en muchas ocasiones que una separación en pasajes pueda provocar que un documento no se considere relevante. No obstante hay que indicar que estos modelos incrementan la complejidad del proceso de búsqueda, sobre todo el modelo *arbitrary-passages*, que es el modelo más complejo de los evaluados.
- El modelo basado en la segmentación en párrafos obtiene unos buenos resultados en una colección tan estructurada como la

del Federal Register con párrafos de considerable tamaño, no obstante sus resultados son sensiblemente peores cuando los párrafos son más pequeños y en general los documentos son menos estructurados como es el caso de la colección LATimes.

- El modelo de *Pages*, soluciona parcialmente el problema que tiene el modelo basado en párrafos, al generar un conjunto homogéneo de pasajes. Así el modelo *Pages*, teniendo en cuenta que es menos complejo que los modelos de ventana con solapamiento, obtiene unos resultados más que aceptables en ambas colecciones.
- El modelo que en general obtiene peores resultados es el modelo de RP semántico *TextTiling*. Se demuestra que este algoritmo de segmentación no consigue una generación de pasajes con contenido que permita mejorar el proceso de RI.
- Como conclusión final cabe volver a citar el buen comportamiento del sistema IR-n en ambas pruebas. Supera a todos los modelos de RI evaluados excepto el *Arbitrary passages* y *Sliding windows* en la colección Federal Register a nivel de AvgP. No obstante, cabe destacar que el sistema IR-n obtiene los mejores resultados a nivel de precisión a los cinco documentos recuperados, genera pasajes con entidad sintáctica (no lo hacen así estos dos modelos) y además su complejidad es menor.

5.6 Conclusiones del capítulo

En este capítulo se ha presentado de forma detallada el diseño y realización de los procesos de entrenamiento y evaluación del sistema IR-n.

En primer lugar, se han descrito los principales modelos de evaluación de sistemas de RI y se ha justificado la elección del método de las colecciones de test aplicado en este trabajo.

En segundo lugar, se ha detallado el proceso de entrenamiento realizado. Este entrenamiento ha permitido parametrizar el sistema IR-n de forma que permita optimizar sus resultados.

En tercer lugar, se ha detallado el proceso de evaluación acometido para medir el rendimiento final del sistema. Esta evaluación

ha consistido en valorar los resultados obtenidos por el sistema IR-n en las conferencias del CLEF-2002 en las que ha participado. Esta participación ha permitido evaluar de forma independiente y externa el sistema, y comparar sus resultados con los principales modelos de RI existentes actualmente. Posteriormente, se ha realizado un análisis de los resultados obtenidos. Éstos demuestran que el sistema IR-n tiene un rendimiento en la tarea de RI sensiblemente superior a la media de todos los sistemas participantes en la conferencia CLEF-2002. Además, cabe destacar que es el mejor sistema cuando se evalúa la precisión a los cinco documentos recuperados.

Para finalizar, se ha comparado el rendimiento del sistema IR-n frente a los modelos de RP más conocidos. Esta evaluación permite afirmar que el sistema IR-n tiene un comportamiento por superior o similar que la mayoría de modelos de RP propuestos.

Una vez presentados y analizados los resultados que el sistema IR-n ha obtenido en tareas de RI, en el siguiente capítulo se va a estudiar cómo se puede utilizar el sistema IR-n en otras tareas muy relacionadas con la RI, como son la BR y la SID.

6. Evaluación del sistema IR-n en otras tareas

En este capítulo se presenta una visión conjunta de las tareas de entrenamiento y evaluación del sistema IR-n en tareas de Búsqueda de Respuestas (BR) y de Selección Interactiva de Documentos (SID).

En primer lugar, se detallarán los objetivos a cumplir y la problemática de cada una de dichas tareas y se estudiará cómo se puede utilizar el sistema IR-n para facilitar la realización de las mismas.

En segundo lugar, se indicará el proceso de entrenamiento que se ha realizado con el sistema IR-n para adecuarlo a estas tareas. Además, se mostrarán los resultados de la evaluación del sistema IR-n obtenido en las conferencias TREC y CLEF, en las tareas objeto de estudio de este capítulo. Finalmente, se mostrarán las conclusiones obtenidas en la aplicación del sistema IR-n en ambas tareas.

6.1 El sistema IR-n en la tarea de Búsqueda de Respuestas

La BR se puede definir como aquella tarea automática, realizada por ordenadores, que tiene como finalidad encontrar respuestas concretas a necesidades precisas de información formuladas por los usuarios. Como ya se indicó en el capítulo 1, los sistemas de BR son especialmente útiles en situaciones en las que el usuario final necesita conocer un dato muy específico y no dispone de tiempo –o no necesita– leer toda la documentación referente al tema de la búsqueda para solucionar su problema.

Los primeros sistemas de BR se basaron en el uso de técnicas de RI adaptadas a la tarea de BR. Estos sistemas son considerablemente eficaces en la detección de documentos y/o fragmentos de los mismos susceptibles de contener la respuesta buscada. Sin embargo, su rendimiento se reduce drásticamente cuando el sistema debe localizar y extraer la respuesta concreta, o bien, un extracto de texto muy reducido que la contenga. Esto es debido a que si un sistema de BR quiere contestar satisfactoriamente una pregunta de un usuario, necesita entender, hasta unos niveles mínimos, tanto la pregunta como la colección de textos donde puede hallarse la respuesta.

Este problema se solucionó mediante la aplicación de técnicas de procesamiento del lenguaje natural (PLN). No obstante, su elevado coste computacional aconsejó emplearlas en las etapas finales de localización y extracción de la respuesta. Por ello, con la finalidad de poder compatibilizar y aprovechar las ventajas de las técnicas de RI y PLN, la gran mayoría de sistemas las emplean en dos fases. En una primera fase se utilizan técnicas de RI para seleccionar aquellos documentos o pasajes susceptibles de contener la respuesta buscada. Posteriormente, en una segunda fase se aplican técnicas de PLN únicamente sobre estos textos relevantes. Con ello se permite la localización de la respuesta correcta, reduciendo sensiblemente el coste de aplicación de las técnicas de PLN, al utilizarlas sólo sobre la preselección de textos realizada.

Éste es el enfoque utilizado por el sistema de BR denominado *SEMQA* (Vicedo, 2002) presentado en las conferencias TREC-9. Para localizar la respuesta a una pregunta determinada, el sistema *SEMQA* solamente consideraba los cincuenta documentos considerados más relevantes por un sistema de RI. Este esquema de trabajo puede verse en la figura 6.1.

La RP en la que se basa el modelo IR-n, ofrece una serie de ventajas dentro de esta tarea de preselección de documentos para un sistema de BR. Fundamentalmente éstas son dos:

- La recuperación de un pasaje relevante en lugar del documento completo, con lo que se disminuye la cantidad de información

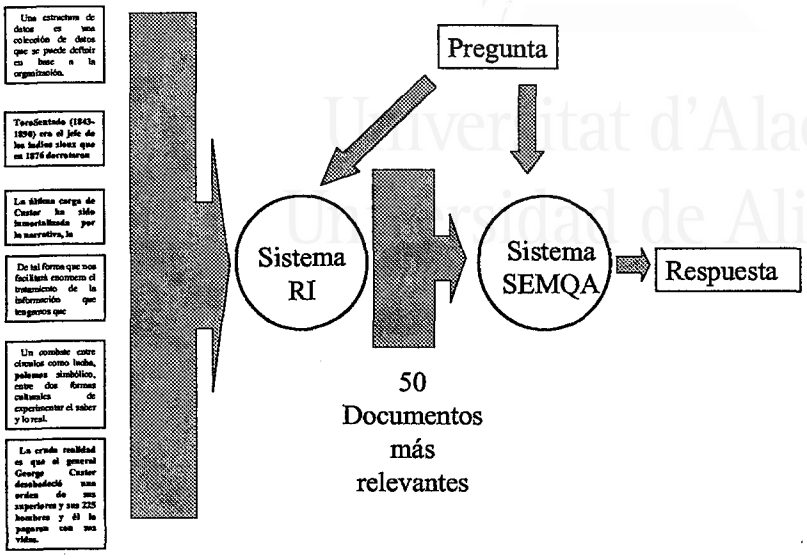


Figura 6.1. Sistema SEMQA en el TREC-9

a procesar por el sistema de BR. Esta circunstancia permite disminuir el tiempo global del proceso de BR.

- Los sistemas de RP permiten detectar pasajes cortos de gran relevancia que forman parte de un documento que, en su conjunto puede no serlo. Estos documentos, estudiados por un sistema de RI basado en el análisis del documento completo, pueden considerarse no relevantes y, en consecuencia, se descartarían aún conteniendo la respuesta correcta.

Además, el modelo de RI que propone el sistema IR-n, es decir el uso de frases completas como unidad, permite mantener la estructura sintáctica en los pasajes que selecciona, con lo cual, se facilita la aplicación de las técnicas de PLN que suelen utilizar los sistemas de BR en dichos pasajes seleccionados.

El objetivo del estudio de la aplicación del sistema IR-n a tareas de BR, consiste en parametrizar el mismo de forma que pueda

proporcionar a un sistema de BR una serie de pasajes relevantes para cada pregunta de forma que:

- Se maximice el número de preguntas para las que se suministran documentos o fragmentos de los mismos que contienen respuesta a dicha pregunta.
- Se minimice la cantidad de información que el sistema de BR debe procesar.

Para cumplir estos objetivos se plantea realizar un estudio de la aplicación del sistema IR-n sobre los datos de la conferencia TREC-9, para posteriormente participar de forma conjunta con el sistema *SEMQA* en la convocatoria TREC-10. Este estudio se detalla en los siguientes apartados.

6.2 La tarea de BR en la conferencia TREC-9

Las conferencias del TREC incluyeron una tarea de BR en el año 1999. El objetivo de esta tarea era que los sistemas participantes seleccionaran de forma automática, a partir de una colección de documentos, pequeños fragmentos de texto que contuviesen la respuesta a una pregunta determinada.

De forma muy similar a las tareas del CLEF, comentadas en el capítulo anterior, en las conferencias TREC y dentro de ellas, en la tarea de BR, se suministra a los participantes una colección de documentos de gran tamaño y un conjunto de preguntas cuya respuesta ha de extraerse a partir de dichos documentos.

Los sistemas participantes deben devolver las respuestas concretas a cada una de las preguntas en un tiempo determinado, (en este caso 1 semana). Este tiempo es mucho menor que el que se utiliza en las conferencias CLEF, que suele estar cercano a los 30 días.

Una vez los sistemas participantes han obtenido y enviado sus respuestas, la organización las evalúa y calcula los resultados de cada uno de los sistemas. Los elementos principales de este proceso son los siguientes:

- La colección de documentos

- La colección de preguntas
- Los resultados obtenidos
- Criterios de relevancia
- Medidas de evaluación

Estos elementos se describen a continuación.

6.2.1 La colección de documentos

La colección de documentos utilizada en la conferencia del TREC-9 está formada por un total de 978.952 documentos de las siguientes colecciones:

- **Associated Press Newswire.** Esta colección está formada por un conjunto de noticias aparecidas tanto en prensa escrita como en emisiones radiofónicas en los años 1988, 1989 y 1990 en medios de difusión relacionados con Associated Press.
- **Wall Street Journal.** Colección formada por un conjunto de noticias publicadas por el periódico financiero The Wall Street Journal. Incluye material de los años 1987, 1988, 1989, 1990 y 1991.
- **San Jose Mercury News.** Corresponde a un conjunto de noticias del año 1991 publicadas por el diario San Jose Mercury News.
- **Financial Times.** Colección que recoge las noticias publicadas por este periódico financiero en los años 1991, 1992, 1993 y 1994.
- **Los Angeles Times.** Esta colección está formada por aproximadamente el 40% de los artículos publicados en el periódico Los Angeles Times en el periodo desde el 1 de enero de 1989 al 31 de diciembre de 1990.
- **Foreign Broadcast Information Service.** Este organismo se encarga de recopilar y traducir, para el gobierno de los Estados Unidos, información de carácter político, económico, técnico y militar procedente de los medios de comunicación de todo el mundo. El acceso a la información del FBIS está limitado a las agencias estatales de dicho país y a sus contratistas. Esta colección está compuesta por un conjunto de noticias recogidas durante 1994.

Las principales características de estas colecciones, referentes a tamaño de la colección en bytes, número de documentos, así como la media y mediana de términos por documento se muestran en la tabla 6.1.

Colección de documentos	Tamaño (MB)	Num. Docs.	Mediana Term. /Doc.	Media Term. /Doc.
Associated Press newswire, 1988	254	84.678	446	473,9
Associated Press newswire, 1989	237	79.919	438	468,7
Associated Press newswire, 1990	237	78.321	451	478,4
Wall Street Journal, 1987-1989	267	98.732	245	434,0
San Jose Mercury News, 1991	287	90.257	379	453,0
The Financial Times, 1991-1994	564	210.158	316	412,7
The L.A. Times	475	131.896	351	526,5
Foreign Broadcast Information Service	470	130.471	322	543,6

Tabla 6.1. Colecciones de documentos de las edición TREC-9

6.2.2 La colección de preguntas

La colección de preguntas fue generada a partir de preguntas reales efectuadas por usuarios de Internet al sistema Encarta de Microsoft[1] y al motor de búsqueda de documentos Excite[2]. Estas preguntas son de tipo corto, es decir formadas por pocas palabras y muy concretas en cuanto a la información solicitada. Ejemplos de preguntas utilizados son "What is the tallest mountain?" o "Where is Glasgow?". Otro aspecto fundamental a destacar es que las preguntas son de dominio abierto.

El número de preguntas que forman parte de la colección es de 693. No obstante cabe indicar que 193 eran variaciones sintácticas de otras preguntas, con lo que realmente el número de preguntas originales era de 500. De dicho conjunto había 8 que no tenían o no se localizó respuesta en la colección. En los experimentos que hemos realizado solamente se han utilizado las 492 preguntas con respuesta en la colección de documentos.

[1] http://encarta.msn.com
[2] http://www.excite.com

6.2.3 Ficheros de resultados

Para cada pregunta, los sistemas de BR podían devolver un máximo de 5 respuestas posibles, debiendo indicar además, el documento de la colección del que se obtuvo cada respuesta.

Se han definido dos posibles longitudes de respuesta, 50 y 250 caracteres como máximo.

Esta información se almacenaba en los ficheros de resultados que debían ser enviados a la organización del TREC dentro de los plazos fijados. El formato de estos ficheros es el indicado en (37).

(37) 241 Q0 EFE19940407-03243 0 Nicole Kidman
241 Q0 EFE19940406-02595 1 Marlon Brando
241 Q0 EFE19940405-01875 2 Penelope Cruz
241 Q0 EFE19940406-02364 3 Lucas Domínguez
241 Q0 EFE19940519-11553 4 Antonio Canales

Donde cada campo se utiliza de la siguiente forma:

1. Número de la pregunta.
2. Campo no utilizado.
3. Número de documento que contiene la respuesta.
4. Posición. Debe contener valores entre 0 y 4, indicando 0 la posición de la respuesta más relevante y 4 la menos relevante de las seleccionados.
5. Cadena que contiene la respuesta.

6.2.4 Criterios de relevancia

La corrección de las respuestas suministradas por los sistemas se realiza de forma manual. Para cada respuesta un grupo de asesores determina de forma individual la validez de la misma.

La opinión de cada uno de ellos puede tener tres posibles juicios:

- **Incorrecta.** La cadena no contiene la respuesta correcta o no está completa.
- **No soportada.** La cadena contiene la respuesta, pero el documento que la contiene no justifica la respuesta.
- **Correcta.** La cadena contiene exactamente la respuesta.

6.2.5 Medidas de evaluación

Una vez evaluada la corrección de cada una de las respuestas, es necesario disponer de una medida que cuantifique el rendimiento general del sistema. Para ello, se emplea la *media recíproca* (*mean reciprocal rank* - MRR). Esta medida se calcula de la siguiente forma. Cada pregunta se puntúa de forma individual con el valor inverso de la posición en la que se encuentra la primera respuesta correcta, o cero si no aparece la respuesta correcta entre las 5 respuestas devueltas por el sistema. La media recíproca computa la media de los valores individuales alcanzados para cada pregunta de la colección de test según la siguiente expresión:

$$MRR = (\sum_{i=1}^{Q} \frac{1}{far(i)})/Q \tag{6.1}$$

donde Q corresponde al número de preguntas de test y *far(i)* indica la posición de la primera respuesta correcta para la pregunta **i**. El valor de *(1/far(i))* será *cero* si no se ha encontrado la respuesta.

Un ejemplo de cálculo de esta medida es el siguiente. Dada una prueba de tres preguntas, de las cuales para la primera no se ha encontrado la respuesta correcta y para la segunda y tercera dicha respuesta se halla en las posiciones primera y cuarta respectivamente. La MRR es igual a $(0 + 1/1 + 1/4)/3$, o sea 0,416.

Debido a la existencia de diferentes niveles de corrección en las respuestas, correctas o injustificadas, la media recíproca puede ser diferente en función de cómo se valoren las respuestas consideradas "injustificadas". Por ello, el rendimiento de los sistemas se valora en función de dos medidas:

- **Valor estricto (strict score)**. La media recíproca estricta se calcula teniendo en cuenta únicamente las respuestas evaluadas como "correctas". Las restantes respuestas se consideran todas "incorrectas".
- **Valor permisivo (lenient score)**. En este caso, el cálculo de la media recíproca se realiza considerando también como "correctas" aquellas respuestas catalogadas como "injustificadas".

6.3 Entrenamiento del sistema IR-n en tareas de BR

El principal objetivo a cumplir por el sistema IR-n es seleccionar una serie de fragmentos de texto que sean relevantes a una pregunta concreta. Estos fragmentos deben contener el mayor número de respuestas y tener el menor tamaño posible.

Por ello, en este entrenamiento se determinará la forma en la que se debe utilizar el sistema IR-n para conseguir dicho objetivo dentro de la tarea de BR.

Los parámetros a ajustar en este entrenamiento son los siguientes:

1. Tamaño del pasaje a utilizar.
2. Medida de similitud a utilizar.
3. Si permite obtener mejores resultados la utilización de técnicas de expansión de la pregunta.

Además, otro elemento a calcular es el número de pasajes que se deben suministrar a un sistema de BR, con el objeto de maximizar el número de respuestas que incluyan dichos pasajes y minimizar la cantidad de información que contengan.

Para cumplir estos objetivos se ha planteado realizar una serie de pruebas con los modelos del sistema IR-n expuestos en el capítulo 4 para adaptarlos al proceso de BR.

La colección de prueba utilizada ha sido la colección de documentos y preguntas de las conferencias TREC-9 descritas en el apartado anterior. Con el objeto de realizar el mayor número de

pruebas significativas, algunos experimentos sólo han utilizado un subconjunto más reducido de preguntas.

Además, hay que destacar que se ha utilizado una herramienta suministrada por la organización de la conferencia TREC, que permite conocer de forma automática si un fragmento de texto contiene la respuesta concreta a una pregunta. La evaluación de todas las pruebas de BR descritas en este capítulo se ha realizado utilizando dicha herramienta.

6.3.1 Experimentos realizados

La determinación de las características óptimas de funcionamiento del sistema IR-n se realizaron mediante una serie de experimentos, agrupados bajo los siguientes epígrafes:

- **Experimento 1. Determinación de la medida de similitud a aplicar.** Se realizarán pruebas utilizando las medidas $IR - n_{base}$ e $IR - n_{prox}$ con el objeto de comprobar cuál es la que permite obtener mejores resultados.
- **Experimento 2. Aplicación de medidas de expansión de la pregunta.** Se realizarán pruebas de uso de las técnicas de expansión de la pregunta incorporadas al sistema IR-n, para verificar si permite obtener mejores resultados que sin aplicar dichas técnicas.
- **Experimento 3. Comprobación con el conjunto de preguntas completo.** Los experimentos 1 y 2 se han realizado con las primeras 100 preguntas de toda la colección del TREC-9. Este experimento final comprobará que los resultados obtenidos con dichas 100 preguntas son significativos. Para ello, se repetirán las pruebas que mejores resultados han obtenido, pero utilizando el juego de preguntas completo. Finalmente se contrastarán los resultados.

6.3.2 Visualización de resultados

Se ha tomado como base para la comparación de los sistemas dos elementos:

- La medida propuesta por la organización, la MRR. Esta medida favorece los sistemas que recuperan las respuestas en los primeros lugares.
- El número de preguntas para las que al menos se ha localizado un pasaje que contiene la respuesta.

Dado que se van a realizar pruebas para diferentes tamaños de pasajes, los resultados se mostrarán en forma de tabla. En las columnas se mostrará el tamaño de pasaje utilizado. En las filas se visualizará el numero de pasajes recuperados. En la intersección de las mismas se indicará el número de preguntas para las que al menos se ha recuperado un pasaje que contiene la respuesta. La última fila mostrará el valor de la MRR para cada una de las pruebas con diferentes tamaños de pasajes.

Todas las pruebas se han realizado sobre 300 pasajes recuperados para cada pregunta. Esto se debe a que hay un tiempo máximo, fijado por la organización para la realización de la tarea, y se ha comprobado que el sistema de BR utilizado puede asumir el proceso de esta cantidad de pasajes en dicho tiempo.

La tabla 6.2 muestra un ejemplo de esta forma de visualizar los resultados. En las columnas aparecen los tamaños de pasajes para los que se ha realizado alguna prueba (5, 10, 15, 20, 25 y 30). En las filas se indica el número de pasajes que se evalúan.

Así por ejemplo en la tercera columna (tamaño = 10) se muestra que el número de preguntas para las que al menos se ha recuperado un pasaje que contiene la respuesta son a los 5 pasajes 71, a los 10 pasajes 79 y así sucesivamente. En la última fila se muestran los valores de la MRR cuando se han recuperado 300 pasajes para cada uno de los tamaños.

6.3.3 Experimento 1. Determinación de la medida de similitud a aplicar

Objetivo. El objetivo de este experimento es determinar la medida de similitud más adecuada que debe utilizar el sistema IR-n para procesos de BR, así como el número de pasajes y el tamaño de los mismos.

Número de pasajes (P)	Tamaño del pasaje (T)					
	5	10	15	20	25	30
5	62	71	73	73	78	79
10	70	79	81	82	82	85
20	80	82	84	86	88	89
30	84	89	88	87	90	89
50	84	94	94	94	95	95
100	91	96	96	96	98	98
200	92	97	97	97	98	98
300	93	97	97	97	99	99
MRR	0,470	0,538	0,572	0,578	0,612	0,617

Tabla 6.2. Ejemplo de tabla de resultados

Descripción del experimento. Se ha aplicado a las dos principales medidas descritas en el capítulo 4 ($IR-n_{base}$ e $IR-n_{prox}$) del sistema IR-n. En este experimento se han utilizado las primeras 100 preguntas de la colección TREC-9.

Resultados obtenidos. Los resultados de aplicar los modelos $IR - n_{base}$ e $IR - n_{prox}$ se pueden ver en las tablas 6.3 y 6.4 respectivamente.

Número de pasajes (P)	Tamaño del pasaje (T)					
	5	10	15	20	25	30
5	62	71	73	73	78	79
10	70	79	81	82	82	85
20	80	82	84	86	88	89
30	84	89	88	87	90	89
50	84	94	94	94	95	95
100	91	96	96	96	98	98
200	92	97	97	97	98	98
300	93	97	97	97	99	99
MRR	0,470	0,538	0,572	0,578	0,612	0,617

Tabla 6.3. Resultados en BR el modelo $IR - n_{base}$

Número de pasajes (P)	Tamaño del pasaje (T)					
	5	10	15	20	25	30
5	63	68	75	76	78	82
10	71	79	80	81	82	84
20	81	86	86	87	88	91
30	84	91	90	89	90	92
50	84	94	94	94	95	95
100	91	96	96	96	97	98
200	93	97	97	97	97	98
300	93	97	97	97	98	99
MRR	0,493	0,542	0,588	0,597	0,606	0,622

Tabla 6.4. Resultados en BR el modelo $IR - n_{prox}$

Conclusiones del experimento. Las conclusiones a extraer son varias, tanto a nivel de comparativa de los modelos, como de la eficacia del sistema IR-n en la tarea de BR.

1. Al incrementar el tamaño de los pasajes, en general se mejoran los resultados. No obstante hay que destacar, que en algunos casos al incrementar el tamaño del pasaje no se mejoran los resultados. Esto se debe a que el tamaño del pasaje también interviene en el cálculo de la relevancia de los documentos, y se ha podido comprobar en el capítulo 5, que a medida que los tamaños de pasaje utilizados se incrementan a partir de determinado nivel se empeoran los resultados.
2. En la mayoría de los tamaños de pasajes probados, el modelo que mejores resultados obtiene es el sistema $IR - n_{prox}$. Esto es consecuencia del tipo de preguntas utilizadas en el TREC, ya que éstas suelen estar formadas por pocos términos muy relacionados. El modelo $IR - n_{prox}$ favorece los documentos que contienen los términos consecutivos de la pregunta en la misma frase.
   Los resultados comparados del uso de las medidas $IR - n_{base}$ y $IR - n_{prox}$ a nivel de MRR se muestran en la tabla 6.5. En la última fila se muestra el porcentaje de mejora, que en algunos casos está cercano al 5%.

Modelo Aplicado	Tamaño del pasaje (T)					
	5	10	15	20	25	30
Base	0,470	0,538	0,572	0,578	**0,612**	0,617
Prox.	**0,493**	**0,542**	**0,588**	**0,597**	0,606	**0,622**
Incremento	+4,9%	+0,7%	+2,8%	+3,3%	-1%	+0,8%

Tabla 6.5. Comparativa de los valores de MRR utilizando varias medidas de similitud del sistema IR-n

3. El crecimiento a nivel de preguntas con respuesta no es proporcional al tamaño del pasaje utilizado. Así, en el modelo $IR - n_{prox}$ el crecimiento de la MRR a medida que aumenta el tamaño de los pasajes no es lineal, sino que tiene un crecimiento notable cuando se pasa de 5 a 10 pasajes, y de 10 a 15, mientras que en el resto de incrementos del tamaño del pasaje, los valores de la MRR aumentan ligeramente.
   Esto implica que a partir de tamaños de pasajes de 15 frases, se incrementa notablemente la cantidad de información a procesar sin obtener como recompensa un aumento proporcional en la MRR. Estos valores pueden verse en la tabla 6.6. En la misma se puede ver que al pasar de pasajes de 5 a 10 frases, y de 10 a 15 frases el incremento es del 9,9% y del 8,5% respectivamente, mientras que el resto son poco significativos (sobre el 2%).

Modelo Aplicado	Tamaño del pasaje (T)					
	5	10	15	20	25	30
Prox.	0,493	0,542	0,588	0,597	0,606	0,622
Incremento	0	+9,9%	+8,5%	+1,6%	+1,5%	+2,6%

Tabla 6.6. Incrementos de la MRR con el incremento del tamaño del pasaje

4. Por último hay que destacar la elevada eficacia del sistema IR-n en esta tarea. Esto se demuestra por el alto porcentaje de documentos para los que al menos se ha localizado un pasaje que contiene la solución. Utilizando pasajes de 10 frases, se

localizan el 96% de respuestas tan sólo considerando los 100 primeros pasajes.

6.3.4 Experimento 2. Aplicación del modelo de expansión de la pregunta

Objetivo. Determinar si al aplicar las técnicas de expansión de la pregunta incorporadas al sistema IR-n, se obtienen mejores resultados que cuando no se utilizan dichas técnicas.

Descripción del experimento. Se ha aplicado el modelo de expansión de la pregunta definido en el capítulo 4 (técnicas de análisis local), utilizando los mismos tamaños de pasajes (de 5 a 30 frases).

Resultados obtenidos. Los resultados de este experimento pueden verse en la tabla 6.7.

Número de pasajes (P)	Tamaño del pasaje (T)					
	5	10	15	20	25	30
5	62	68	71	71	71	74
10	70	76	79	79	80	82
20	80	82	85	85	86	88
30	81	86	87	87	88	89
50	83	88	88	88	91	93
100	89	90	91	91	93	94
200	94	95	97	97	97	97
300	96	96	97	97	97	98
MRR	0,493	0,519	0,553	0,567	0,564	0,595

Tabla 6.7. Resultados en BR el modelo IR-n con expansión de la pregunta

Conclusiones del experimento. La principal conclusión que se extrae es que al utilizar técnicas de expansión de la pregunta no se mejoran los resultados en comparación al modelo que no la emplea (resultados descritos en la tabla 6.4).

Las técnicas de expansión de la pregunta permiten localizar documentos que no contengan exactamente los términos de la pregunta, con lo que se incrementa la posibilidad de localizar un número mayor de documentos relevantes. No obstante también introducen cierto "ruido" en las preguntas, lo cual, puede provocar que ciertos documentos relevantes ocupen posiciones posteriores en lista de relevancia. Esto produce una disminución en los valores de MRR.

Otro aspecto a tener en cuenta, es que en estas pruebas sólo se está valorando el hecho de localizar al menos un pasaje que contenga una respuesta. Dado el gran tamaño de la colección del TREC, puede no ser necesario utilizar estas técnicas de expansión.

De los resultados obtenidos se concluye, que no es conveniente emplear en el sistema IR-n dichas técnicas de expansión de la pregunta cuando se utiliza en tareas de BR y la colección de documentos es de tamaño considerable.

6.3.5 Experimento 3. Comprobación con la colección completa de documentos

Objetivo. El objetivo de este experimento es doble. En primer lugar se pretende valorar los resultados obtenidos por el sistema $IR - n_{prox}$ utilizando la colección completa de preguntas del TREC-9. En segundo lugar, esto permitirá comprobar si los dos experimentos anteriores, realizados con las 100 primeras preguntas, ha sido significativo.

Descripción del experimento. Se ha aplicado el modelo de $IR - n_{prox}$ a pasajes de 10 y 15 frases, que son los que en el experimento anterior han obtenido proporcionalmente los mejores resultados. En este experimento se han utilizado las 492 preguntas de la colección del TREC-9 que tenían solución en la colección de documentos. Los resultados han sido comparados con los obtenidos en la misma prueba que empleaban únicamente las 100 primeras preguntas.

Resultados obtenidos En la tabla 6.8 se pueden ver los resultados obtenidos utilizando toda la colección completa.

Num Pasajes	Tamaño del pasaje (T)	
Recuperados	10	15
5	348	362
10	390	398
20	421	423
30	436	438
50	449	454
100	460	462
200	468	469
300	472	473
MRR	0,554	0,569

Tabla 6.8. Resultados del modelo IR-n utilizando la colección completa de preguntas.

Conclusiones del experimento. La primera conclusión que se extrae de los resultados es la alta eficacia del sistema IR-n en este tipo de tareas. Utilizando los 200 primeros pasajes se llega a obtener más de un 95% de preguntas para las que al menos se ha localizado un pasaje que contiene la respuesta.

En segundo lugar, cabe destacar que los resultados son muy similares a los obtenidos en el experimento 1. Esto se puede verificar observando las tablas 6.9 y 6.10. En éstas se comparan los resultados obtenidos sobre las cien primeras preguntas (experimento 1) y sobre la colección completa (experimento 3) utilizando pasajes de 10 y 15 frases. En ambos casos se indican los porcentajes de preguntas para las que se ha localizado un documento que contiene la respuesta. En la última columna se muestra la diferencia porcentual entre ambas. Como puede observarse los resultados son muy similares utilizando ambos conjuntos de preguntas. Las diferencias porcentuales en la mayoría de los casos no superan el 2%.

	Sobre 100	Sobre 492	% Δ
5	68	70,7%	+4
10	79	79,2%	+0,3
20	86	85,6%	+0,5
30	91	88,6%	-2,3
50	94	91,3%	-2,9
100	96	93,5%	-2,6
200	97	95,1%	-2
300	97	95,9%	-1,1
MRR	0,542	0,554	+2,2

Tabla 6.9. Comparativa de colecciones de pruebas utilizando pasajes de 10 frases

	Sobre 100	Sobre 492	% Δ
5	75	73,6%	-1,9
10	80	80,1%	+0,1
20	86	86%	0
30	90	89%	-1,1
50	94	92,3%	-1,8
100	96	93,9%	-2,2
200	97	95,3%	-1,8
300	97	96,1%	-0,9
MRR	0,588	0,569	-3,2

Tabla 6.10. Comparativa de colecciones de pruebas utilizando pasajes de 15 frases

Esta comparativa puede verse de forma gráfica en las figuras 6.2 y 6.3. En estas gráficas se puede observar que las líneas que representan los resultados obtenidos con cien preguntas y con la colección completa son muy similares. Por ello, se valora como significativa la experimentación sobre las 100 preguntas realizada en los experimentos 1 y 2.

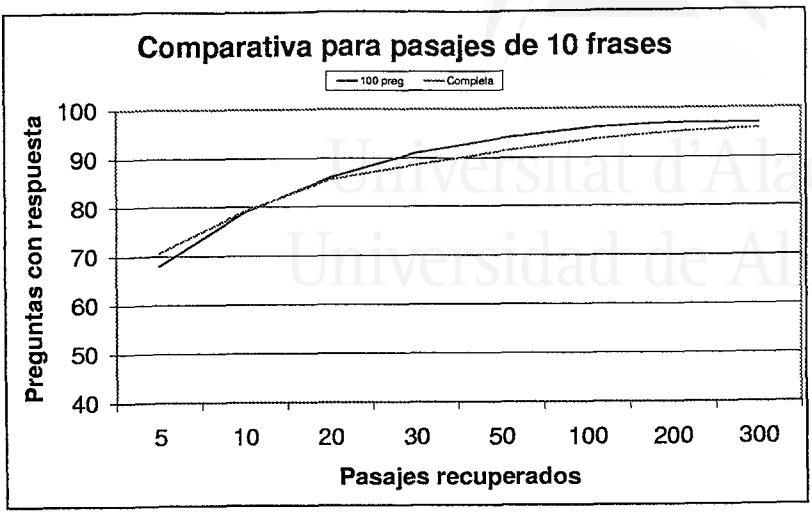


Figura 6.2. Comparativa de resultados utilizando colecciones de 100 preguntas y completa. Pasajes de 10 frases

6.4 Evaluación del sistema IR-n en tareas de BR. Conferencia TREC-10

La evaluación final del sistema presentado en este trabajo se ha realizado empleando la colección de documentos y preguntas de la tarea de BR de la conferencia TREC-10. La evaluación se ha efectuado en base a dos estudios de resultados. En primer lugar, se compara el rendimiento del sistema IR-n frente al modelo del coseno, también empleado como sistema base en el capítulo 5. En segundo lugar, se muestran los resultados oficiales que obtuvo el sistema IR-n en la tarea de BR de la conferencia TREC-10, en la que participó de forma conjunta con el sistema SEMQA de BR.

6.4.1 Descripción de la tarea

Las modificaciones más importantes introducidas en esta edición con respecto a la del año anterior son las siguientes:

1. La longitud máxima de las respuestas se limita a 50 caracteres.

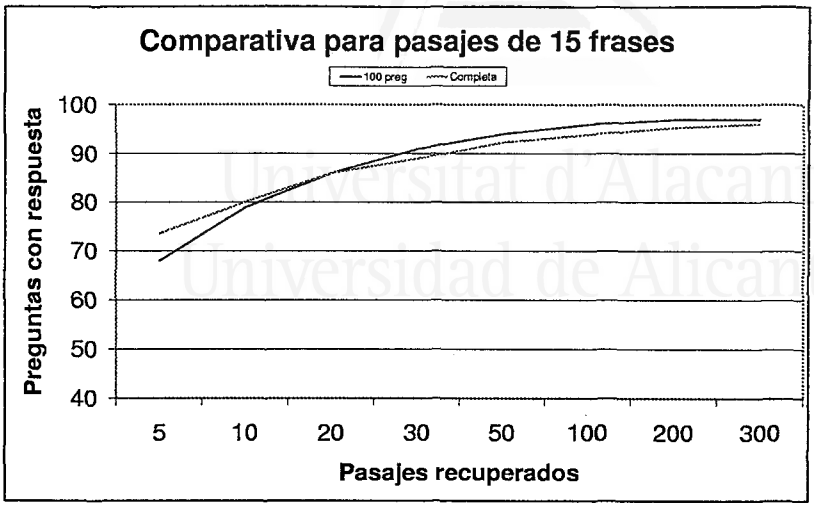


Figura 6.3. Comparativa de resultados utilizando colecciones de 100 preguntas y completa. Pasajes de 15 frases

2. No se garantiza que la colección de documentos contenga respuestas a todas las preguntas propuestas. De esta forma, el sistema puede devolver como respuesta la cadena "NIL" indicativa de que el sistema considera que no existe respuesta en la colección de documentos a la pregunta planteada. La respuesta "NIL" se puntúa de la misma forma que las restantes respuestas. Esta respuesta se contabiliza como "correcta" cuando no se conoce que exista, en la colección, la respuesta a una pregunta.

El conjunto de preguntas de test utilizado para esta tarea se construyó a partir de preguntas reales efectuadas a los sistemas MSNSearch de Microsoft[3] y AskJeeves[4]. Estas colecciones iniciales se filtraron de forma manual para extractar aquel subconjunto de preguntas que se ajustaban a la tarea y que presentaban una formulación correcta. De entre ellas, se seleccionaron 500 preguntas que conformaron el conjunto final de test. No se localizó respuesta alguna en la colección documental para 49 de estas preguntas,

[3] http://search.msn.com
[4] http://www.askjeeves.com

por tanto, la respuesta "NIL" se consideró correcta para dichas preguntas.

6.4.2 Evaluación del sistema IR-n frente al modelo del coseno

Este apartado se va a centrar en la comparación de los resultados obtenidos por el sistema IR-n frente al modelo del coseno, utilizando las 451 preguntas de la colección TREC-10 que tienen respuesta. Para ello, se utilizó el modelo del coseno y el sistema IR-n para recuperar los documentos y pasajes respectivamente más relevantes para cada pregunta. Posteriormente, se comprobó si los documentos suministrados por el modelo del coseno y los pasajes seleccionados por el sistema IR-n contenían la respuesta a cada pregunta. Este proceso se realizó de forma automática utilizando la herramienta suministrada por la organización.

Los resultados obtenidos se muestran en la tabla 6.11. En la misma se incluye el número de preguntas para las que se ha recuperado al menos un documento o pasaje que contiene la respuesta, así como la MRR. En el sistema IR-n se ha utilizado la medida $IR-n_{prox}$ con pasajes de 10 y 15 frases. Estos son los parámetros que permitieron obtener mejores resultados en la fase de experimentación. En dicha tabla también se muestran los incrementos que obtiene el sistema IR-n en cada uno de los intervalos y a nivel de MRR.

Como se puede observar se obtienen sensiblemente mejores resultados utilizando el sistema IR-n sobre el modelo del coseno. Cabe destacar fundamentalmente dos aspectos:

1. El modelo IR-n permite seleccionar en las primeras posiciones más documentos relevantes que el modelo del coseno.
2. El modelo IR-n permite determinar no sólo el documento relevante, sino la parte relevante del mismo. Los resultados que se muestran en la tabla son en base a la localización de la respuesta en el documento completo (modelo del coseno) y el pasaje más relevante (modelo IR-n). Así, el modelo IR-n permite obtener mejores resultados y además disminuye notablemente el tamaño de la información a procesar por el sistema de BR.

Modelo	Coseno	Tamaño del pasaje (T)			
		10	% Δ	15	% Δ
5	232	303	+30,6	309	+33,2
10	281	337	+19,9	375	+33,5
20	324	363	+12,0	383	+18,2
30	349	378	+8,3	394	+12,9
50	370	390	+5,4	406	+ 9,7
100	392	403	+2,8	414	+5,6
200	408	412	+1	417	+2,2
300	415	415	+0	417	+ 0,5
MRR	0,357	0,477	+33,6	0,485	+35,9

Tabla 6.11. Resultados en TREC-10

6.4.3 Evaluación oficial del modelo SEMQA + IR-n

En la conferencia TREC-10 se utilizó el sistema IR-n para la selección de los pasajes más relevantes para cada pregunta. El sistema SEMQA realizó la tarea de localizar dentro de dichos pasajes las respuestas a cada pregunta.

Una vez estudiados los resultados del entrenamiento, se decidió que el sistema IR-n seleccionara los 200 pasajes de 15 frases más relevantes para cada pregunta. Esto se debe a que con este tamaño de pasajes se obtuvieron los mejores resultados en proporción al tamaño de los mismos. Además, se consideró que dado el escaso tiempo que se dispone para el proceso de todos los pasajes, el sistema SEMQA podía procesar los 200 pasajes. Este esquema de trabajo puede verse en la figura 6.4.

La tabla 6.12 muestra los resultados obtenidos por cada uno de los sistemas participantes en la tarea principal TREC-10. Esta tabla presenta los sistemas de forma ordenada en función de su rendimiento estricto (MRR strict). En esta tabla se indican el valor de MRR y el porcentaje de preguntas para las que se ha encontrado respuesta.

Como se puede comprobar en dicha tabla el sistema (sistemas SEMQA e IR-n) ocupa la posición doceava de los treinta y seis sistemas que se habían presentado, lo cual puede calificarse de po-

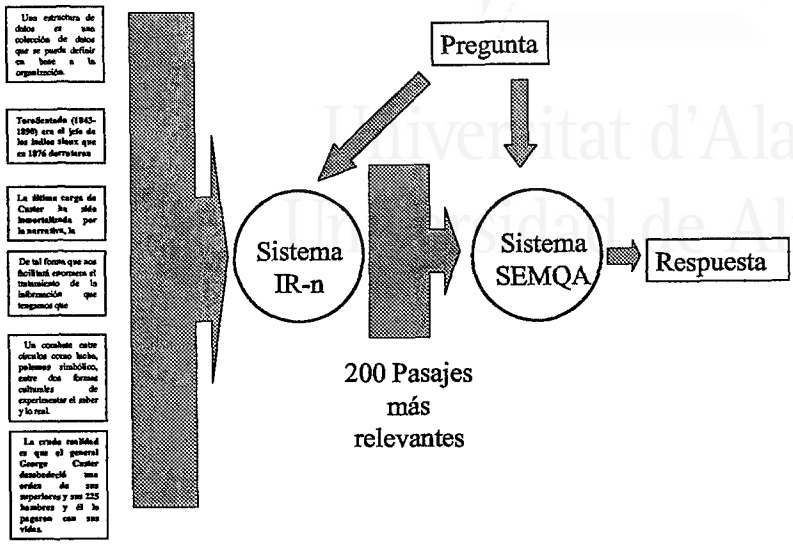


Figura 6.4. Sistema IR-n y SEMQA en el TREC-10

sitivo. Además, se han obtenido resultados superiores a la media de los participantes, esto se puede comprobar en la tabla 6.13.

6.5 Análisis de los resultados obtenidos por el sistema IR-n en tareas de BR

Dados los resultados descritos en el apartado anterior se puede afirmar que el sistema IR-n es adecuado para la preselección de pasajes más relevantes dentro de la tarea de BR.

Tal como se puede observar en la tabla 6.11, el esquema utilizado por el sistema SEMQA en el TREC-9 (seleccionar los cincuenta documentos más relevantes), aplicado sobre las colecciones del TREC-10, hubiera permitido localizar un máximo de 370 respuestas. Sobre la misma colección, el sistema IR-n, seleccionando los 200 pasajes de 15 frases más relevantes, se posibilita localizar un 92% de las respuestas, (417 sobre 451 preguntas con respues-

Comparativa resultados TREC-10 (tarea principal)				
Organización	Strict MRR	Strict %C	Lenient MRR	Lenient %C
1 InsightSoft (Soubbotin y Soubbotin, 2002)	0,676	69,1	0,686	70,1
2 LCC (Harabagiu et al., 2002)	0,570	65,2	0,587	67,7
3 Oracle (Alpha et al., 2002)	0,477	60,8	0,491	62,6
4 USC - ISI (Hovy et al., 2002)	0,435	58,3	0,451	60,2
5 University of Waterloo (Clarke et al., 2002)	0,434	56,9	0,457	59,3
6 Sun Microsystems Lab. (Woods et al., ????)	0,405	55,3	0,418	56,7
7 IBM (Ittycheriah et al., 2002)	0,390	55,7	0,403	56,9
8 IBM (Prager et al., 2002)	0,357	55,3	0,365	57,1
9 Microsoft (Brill et al., 2002)	0,347	50,4	0,437	60,4
10 Queens College, CUNY (Kwok et al., 2002)	0,326	46,3	0,331	47,2
11 POSTECH (Lee et al., 2002)	0,320	43,9	0,335	47,2
12 U.Alicante(SEMQA+IR-n) (Vicedo et al., 2002)	0,300	39,6	0,306	40,4
13 University of Alberta	0,299	38,2	0,311	39,0
14 Korea University	0,294	39,4	0,298	40,0
15 University of Pisa (Attardi et al., 2002)	0,270	32,3	0,271	32,5
16 NTT Com. Science Lab. (Kazawa et al., 2002)	0,228	31,5	0,231	31,9
17 University of Pennsylvania	0,226	33,5	0,235	34,8
18 Syracuse University (Chen et al., 2002)	0,218	32,5	0,230	33,7
19 Tilburg University (Buchholz, 2002)	0,210	27,0	0,234	29,5
20 EC Wise, Inc. (Rennert, 2002)	0,197	28,7	0,204	29,7
21 Université de Montréal (Plamondon et al., 2002)	0,191	34,6	0,197	35,6
22 University of Amsterdam (Monz y de Rijke, 2002)	0,190	26,8	0,203	28,7
23 LIMSI (Ferret et al., 2002)	0,181	26,0	0,192	27,4
24 University Illinois/Champaign (Roth et al., 2002)	0,165	22,0	0,193	25,4
25 Harbin Institute of Technology	0,162	24,6	0,166	25,2
26 KAIST (Oh et al., 2002)	0,152	22,6	0,159	23,6
27 National Taiwan University (Lin y Chen, 2002)	0,145	22,4	0,146	22,8
28 Fudan University (Wu et al., 2002)	0,137	20,7	0,145	22,2
29 KCSL	0,126	22,0	0,131	22,8
30 MITRE	0,125	19,3	0,131	19,7
31 CL Research (Litkowski, 2002)	0,120	19,1	0,130	20,3
32 University of York (Alfonseca et al., 2002)	0,111	16,5	0,121	18,1
33 ITC - IRST (Magnini et al., 2002)	0,105	16,5	0,110	17,1
34 Chinese Academy of Sciences (Wang et al., 2002)	0,100	13,6	0,109	15,4
35 University of Iowa	0,061	12,6	0,064	13,2
36 Conexor Oy	0,003	0,4	0,003	0,4

Tabla 6.12. Comparativa de resultados de los sistemas participantes en la tarea principal TREC-10

ta). Estos datos demuestran la incidencia que ha podido tener el sistema IR-n en los resultados oficiales obtenidos en la edición del TREC-10.

Además de incrementar el número de respuestas, también se puede observar que el sistema IR-n obtiene mejores resultados,

Resultados TREC-10						
	Strict			Lenient		
Prueba	MRR	% Corr.	%Δ	MRR	% Corr.	%Δ
TREC-10	0,234	33,0	0,0	0,246	34,6	0,0
SEMQA+IR-n	0,300	39,6	+20,0	0,306	40,4	+16,8

Tabla 6.13. Resultados de la evaluación

a nivel de preguntas con respuesta, en todos los puntos de corte que se muestran en la misma tabla. Esto quiere decir que, además de disminuir la cantidad de información a procesar por el sistema de BR, el sistema IR-n tiene mejor rendimiento en la tarea de localizar los documentos relevantes que el modelo del coseno.

Por último, cabe recordar que el tipo de pasajes proporcionados por el sistema IR-n, pasajes formados por frases, y por tanto con estructura, facilitan enormemente los procesos que los sistemas de BR deben hacer sobre los mismos para la localización de las respuestas.

6.6 El sistema IR-n en la Selección Interactiva de Documentos

Uno de los problemas de los sistemas de RI que se basan en el documento completo es que son capaces de determinar qué documento es relevante o no, pero no especifican la parte del documento que realmente es relevante a la pregunta. Un documento titulado "Biografía de Felipe II", será relevante con respecto a una pregunta del tipo "Donde nació de Felipe II?", pero sólo una parte de dicho documento será el que hace referencia a dicha pregunta. Así, cuando se le presenta a un usuario un documento considerado relevante por un sistema de este tipo, en ocasiones deberá estudiar de forma completa dicho documento para determinar si es relevante o no. Esto puede incrementar notablemente el tiempo que el usuario debe dedicar a esta tarea.

La ventaja que puede aportar el sistema IR-n en este tipo de tareas es que no sólo es capaz de determinar la relevancia del documento, sino que, además, indica qué pasaje es el más rele-

vante. Esto disminuye notablemente la cantidad de información que el usuario debe leer para determinar la relevancia del documento. Además, como el sistema IR-n devuelve pasajes formados por frases completas, éstos pueden ser fácilmente entendidos por el usuario.

Como ya se ha comentado en el capítulo 3, el potencial usuario de un sistema de RI, no sólo valora el concepto de eficacia medido en base a los resultados obtenidos en cuanto a cobertura y precisión, sino que también tiene en cuenta otros factores como son la interactividad del sistema y la eficiencia del mismo.

En este apartado se detallan las ventajas que ofrece el sistema IR-n para facilitar al usuario la localización de documentos relevantes. Este estudio del sistema IR-n se ha realizado dentro del ámbito de las tareas de Selección Interactiva de Documentos (en adelante SID) de la edición de la conferencia iCLEF-2002 (celebrada bajo los auspicios del CLEF-2002).

6.7 La tarea de Selección de Documentos Interactiva en el iCLEF-2002

La tarea de SDI en el iCLEF-2002 se realiza dentro de las conferencias CLEF. El principal objetivo de esta tarea es suministrar unas pautas comunes que faciliten la comparación y evaluación de sistemas interactivos de RI multilingüe.

Esta tarea se diferencia de las tareas de RI comentadas en el capítulo 5 en que incorpora la idea de interactividad. Mientras que en los experimentos que se realizan en tareas de RI tradicionales el usuario no interviene para determinar qué documentos le son relevantes, en la tarea de SID, el objetivo a cumplir es definir la forma en la que se muestran los documentos considerados relevantes por un sistema de RI, para facilitar al usuario el proceso de determinar si el documento es relevante o no.

La idea de multilingüalidad de la tarea SID se basa en el hecho que la pregunta está realizada en un idioma y la colección de documentos en la que se ha de localizar aquellos que son relevantes se encuentra en otro.

Es un hecho que a un usuario le es más fácil leer y comprender un pequeño fragmento de texto que un documento completo. El estudio que se describe a continuación tiene como objetivo determinar si el uso del sistema IR-n, basado en la localización de los pasajes más relevantes, permite a un usuario la determinación de la relevancia de un documento, de forma más sencilla que lo haría si utilizara un sistema de RI que devuelve el documento completo.

6.7.1 Especificación de la tarea

El diseño de la tarea SID está estructurada en los siguientes pasos:

1. Cada participante debe comparar dos sistemas de SID, uno de ellos es tomado como sistema base.
2. Se definen grupos de usuarios múltiplos de cuatro.
3. Se definen una serie de consultas (en este caso cuatro) en el idioma que conocen los usuarios.
4. Se define la colección a utilizar, que debe estar escrita en un idioma diferente al de las consultas.
5. Para cada una de las preguntas se muestran al usuario de forma ordenada los documentos considerados relevantes, de la forma que el sistema defina.
6. El usuario debe indicar para cada uno de estos documentos si lo considera relevante o no.
7. Cada participante debe enviar la lista de los documentos considerados relevantes por los usuarios.

Por tanto, el objetivo a conseguir es facilitar al usuario en lo posible la tarea de conocer si un documento es relevante o no ante una necesidad de información.

Los elementos principales de este proceso son los siguientes:

1. La colección de documentos
2. La colección de preguntas
3. Ficheros de resultados
4. Criterios de relevancia
5. Medidas de evaluación

6.7.2 La colección de documentos

La tarea SID utiliza un subconjunto de las colecciones del CLEF. Estas colecciones se hallan en diversos idiomas: francés, inglés, alemán y español. La participación del sistema IR-n en esta tarea se ha basado en el uso de preguntas escritas en español para interrogar una colección de documentos en inglés. Las colecciones utilizadas son:

- **Los Angeles Times.** Descrita en el capítulo 5.
- **Noticias de agencia EFE.** Descrita en el capítulo 5.
- **Los Angeles Times, traducida al español.** Dado que ésta es una tarea bilingüe en la que las preguntas se realizan en español y la colección de documentos se halla en inglés, la organización también puso a disposición de los participantes los documentos considerados más relevantes para cada pregunta, traducidos de forma automática del inglés al español utilizando el traductor *systran*[5]. No obstante, se especificaba que no era necesario utilizar estas traducciones.

6.7.3 La colección de preguntas

La colección de preguntas utilizadas es un subconjunto de las utilizadas en la edición CLEF-2001. Esta colección está formada por cuatro preguntas (números 53, 56, 65 y 80) así como una pregunta adicional que podía ser utilizada para el proceso de entrenamiento del sistema (la pregunta número 86).

Las preguntas tienen el mismo formato del CLEF, es decir formadas por la parte de título, descripción y narrativa, pudiendo utilizar cada grupo los elementos de la misma que considere oportunos.

6.7.4 Ficheros de resultados

Una vez efectuados los procesos en los que cada usuario indica los documentos que considera relevantes, los participantes deben entregar a la organización un fichero en formato SGML en el que se

[5] http://www.systransoft.com

indica, para cada uno de los usuarios, qué documentos ha considerado relevantes para cada pregunta. Un ejemplo de este fichero se indica en (38).

(38)
```
<iclef-results site="Alicante">
<experiment searcher="1" topic="1" system="baseline">
<doc docno="AP1994010100001">
<doc docno="AP1994010200003">
<doc docno="AP1994010100023">
</experiment>
<experiment searcher="1" topic="2" system="keywords">
<doc docno="AP19940211-00221">
<doc docno="AP19940409-00111">
<doc docno="AP19940510-00121">
<doc docno="AP19940511-00311">
</experiment>
</iclef-results>
```

En este fichero se puede comprobar cómo para cada grupo de usuario (*experiment-searcher*), pregunta (*topic*) y sistema (*system*) se indican los nombres de los documentos (*docno*) considerados relevantes.

6.7.5 Criterios de relevancia

El juego de preguntas empleado correspondía a la edición del CLEF-2001, para el cual se disponían de los criterios de relevancia, obtenidos por el *pooling* realizado en la edición de dicho año. Únicamente se debe estudiar la relevancia de aquellos documentos que siendo considerados relevantes por algún participante de la edición del iCLEF, no lo hubiesen sido por ninguno de los participantes en las pruebas referentes a la colección LATimes en la edición del CLEF-2001.

Un vez realizada esta tarea, se generan ficheros de relevancia con el mismo formato descrito en el capítulo anterior para la tarea de RI.

6.7.6 Medidas de evaluación

La organización evalúa si los documentos considerados como relevantes por los usuarios lo son realmente, y se calculan los valores de precisión y cobertura de cada sistema. Parece claro que ambas medidas ofrecen un mayor sentido si son evaluadas de forma conjunta. Es por ello que existen medidas que tienen en cuenta los dos valores. La más conocida es la denominada F_α (Rijsbergen, 1979):

$$F_\alpha = \frac{1}{\alpha/P + (1-\alpha)/C} \tag{6.2}$$

Los valores de α por encima del 0,5 dan mayor importancia a la precisión, mientras que los valores por debajo de 0,5 valoran en mayor medida la cobertura. Así, el valor de α dependerá del objetivo a conseguir por el sistema. En las conferencias del iCLEF, dentro de la tarea interactiva (Gonzalo y Oard, 2002), el valor que se dio a α fue de 0,8, ya que se consideró más razonable dejar de encontrar algún documento relevante (menor cobertura), que dar como relevante a algún documento que no lo era (menor precisión).

6.8 Entrenamiento del sistema IR-n en tareas de SID

El objetivo de esta experimentación era el de valorar de qué forma se podía utilizar el sistema IR-n, para facilitar a un usuario la tarea de determinar si un documento es relevante o no dentro de un entorno multilingüe.

Dado que el sistema IR-n no dispone todavía de posibilidades multilingüe, se empleó en este entrenamiento la colección de LATimes traducida de forma automática al español.

De las preguntas sólo se utilizaron la parte de título y descripción, es decir, lo que se denominó preguntas cortas en el capítulo anterior.

Con el objeto de hacer más agradable e intuitivo el trabajo del usuario con el sistema IR-n, se desarrolló un interfaz web. Este interfaz mostraba únicamente los pasajes más relevantes para cada pregunta de forma ordenada en cuanto a su relevancia. De cada documento sólo se mostraba su pasaje más relevante.

Para esta experimentación se seleccionó un usuario y se realizaron pruebas con la pregunta de entrenamiento que puso a disposición la organización del sistema.

Inicialmente, de cada documento se mostraba únicamente el pasaje más relevante. El tamaño del pasaje era de 8 frases, que fue el tamaño utilizado por el sistema IR-n en la edición del CLEF-2002.

Realizadas estas pruebas se detectaron los siguientes problemas:

1. La traducción de la colección LATimes realizada por el *systran* no era buena y producía en ocasiones fragmentos ilegibles. Además provocaba que al no entender cierto fragmento del documento, el usuario desechara dicho documento y pasara al estudio del siguiente.
2. Al usuario le suponía un coste elevado entender el tema del documento.
3. El usuario necesitaba más información de la que aparecía en el pasaje. Esto se debe a que en el sistema IR-n el pasaje más relevante se inicia en la primera frase en la que aparecía un término de la pregunta, con lo cual no se mostraba los antecedentes de dicho pasaje contenidos en frases anteriores.

Para evitar o al menos, minimizar estos problemas, se tomaron las siguientes decisiones:

1. Para evitar que el usuario, al no entender parte del documento, pasara al siguiente directamente, se decidió mostrar las frases en líneas diferentes. Dado que el sistema IR-n realiza una indexación basada en frases, es sencillo visualizarlas de forma

separada. Se comprobó que este hecho facilitaba un mejor entendimiento de los textos.

2. Para facilitar al usuario el conocimiento del tema central del que trataba cada documento, siempre se visualizaba, como primera línea, el título del documento en mayúsculas, aunque no formara parte del pasaje más relevante.
3. Para evitar posibles pérdidas en el contexto del documento, la primera frase que se mostraba era la inmediatamente anterior a la primera del pasaje más relevante.

6.9 Evaluación del sistema IR-n en tareas de SID

El objetivo principal de la tarea SID en el CLEF-2002 es el de confrontar dos sistemas de RI. Por ello, además del sistema IR-n se utilizó el sistema ZPrise[6], un sistema de RI basado en el documento completo.

Para realizar los experimentos intervinieron ocho usuarios, con formación universitaria. En lugar de explicar a los usuarios lo que la comunidad de RI considera como relevante o no relevante, se coordinaron los criterios de relevancia entre ellos mediante la siguiente estrategia: a cada usuario se le dijo que seleccionara los documentos que él considera interesantes para cada una de las preguntas propuestas.

Así, los experimentos se llevaron a cabo de la siguiente forma:

1. Se mostraba la pregunta, formada sólo por la parte de título y descripción de la pregunta original.
2. Para cada una de las preguntas se seleccionaban los 25 documentos más relevantes (en el caso del sistema ZPrise), y los 25 pasajes más relevantes (en el caso del sistema IR-n).
3. Se mostraba al usuario, de forma secuencial, cada uno de estos documentos o pasajes.
4. El usuario marcaba el documento como "de interés" o "no de interés".

[6] ZPrise desarrollado por Darrin Dimmick (NIST). Disponible bajo petición previa en http://www.itl.nist.gov/iaui/894.02/works/papers/zp2/zp2.html

5. No se limitó al usuario el tiempo que debía utilizar para cada pregunta.

Una vez generadas las listas de documentos relevantes para cada pregunta se enviaron a la organización del iCLEF.

La tabla 6.14 muestra los resultados obtenidos por los sistemas IR-n y ZPrise.

Precision media	IR-n	ZPrise
Pregunta 53	0,4601	0,6371
Pregunta 65	0,8098	0,5925
Pregunta 56	0	0
Pregunta 80	0,7653	0,3748
Media	0,5085	0,4011
F_α	0,32	0,22

Tabla 6.14. Precisión media por pregunta

6.10 Análisis de los resultados del sistema IR-n en tareas de SID

Las conclusiones extraídas, se dividen por una parte en lo que son referentes a la eficacia del sistema y por otra, en lo que sería la satisfacción de los usuarios en el uso del sistema propuesto.

El sistema IR-n ha tenido un mejor comportamiento que el sistema ZPrise en los tres niveles estudiados:

1. Precisión media (un 26,8% superior).
2. Cantidad de preguntas con mejor media de precisión individual (en el sistema IR-n se obtuvieron mejores resultados con 2 preguntas, mientras que con ZPrise únicamente se obtuvieron mejores resultados en una).
3. Valor de F_α (un 45,5% superior).

Este mejor comportamiento se debe, fundamentalmente a dos aspectos. Por un lado el sistema IR-n es mejor que el sistema ZPrise en la tarea de ordenar los documentos por su relevancia, con lo que al usuario se le mostraba en los primeros lugares los

documentos más relevantes. Por otro lado, el hecho de mostrar sólo el pasaje más relevante facilitaba y hacía menos tediosa la labor al usuario.

Dentro de las conclusiones obtenidas a partir de entrevistas realizadas a los usuarios cabe indicar que:

1. Los usuarios se han quejado unánimemente de las traducciones automáticas que se les han mostrado. Indican que en ocasiones, más de las esperadas, los documentos eran auténticamente ilegibles.
2. El hecho de que para la pregunta número 56, no existiera ningún documento relevante provocó cierta ansiedad en los usuarios cuya reacción consistió en considerar relevantes algunos documentos que no lo eran, al temer que no habían detectado algún documento relevante.
3. El mostrar el título en mayúsculas facilitó enormemente la comprensión de los artículos.
4. Los usuarios también indicaron que la presentación de frases en líneas diferentes facilitaba la comprensión de algunas traducciones poco legibles.
5. Otro aspecto que los usuarios han indicado como negativo ha sido el que el contenido de las preguntas era excesivamente escueto y en ocasiones no permitía aclarar exactamente el motivo para considerar relevante un documento.

Finalmente dentro de lo que serían las conclusiones generales de la prueba realizada cabe indicar que:

1. Ha sido complicado encontrar y motivar a los usuarios para realizar las pruebas, provocando que se disminuyese el número de documentos a estudiar para cada pregunta (sólo 25). Esto ha disminuido notablemente los valores posibles de cobertura máximo que se podían obtener. No obstante, el rendimiento del sistema IR-n ha sido excelente y, prácticamente, ha permitido a los usuarios detectar un porcentaje muy elevado de los documentos relevantes mostrados.
2. También ha sido positivo comprobar que los usuarios, al tener que verificar sólo un pasaje de 9 frases (8 frases del pasaje más

relevante más la inmediatamente anterior), han completado la tarea de detección de documentos relevantes en un tiempo muy breve. La media ha estado entre 8 y 9 minutos por pregunta.

6.11 Conclusiones del capítulo

En este capítulo se han presentado de forma detallada los experimentos y la evaluación realizada con el sistema IR-n en otras tareas diferentes a la RI. Estas tareas han sido la Búsqueda de Respuestas y la Selección Interactiva de Documentos.

Como principales conclusiones caben destacar las siguientes:

1. El sistema IR-n obtiene muy buenos resultados en la tarea de preseleccionar los pasajes más relevantes para ser utilizados por un sistema de BR.
2. El modelo de similitud $IR-n_{prox}$ se ha mostrado como el más adecuado para la tarea de BR, debido al tipo de las preguntas utilizadas en este tipo de sistemas.
3. Además, el formato de los pasajes proporcionados por el sistema IR-n (con estructura sintáctica), permiten su tratamiento de forma sencilla y eficaz, tanto en tareas de BR como en tareas SID.
4. El sistema IR-n, al ser capaz de detectar el pasaje más relevante de un documento, es una herramienta eficaz dentro de las tareas de BR y SID, ya que permite al sistema de BR o al usuario centrarse en las partes más relevantes de los documentos a estudiar.

Estos hechos se han refrendado mediante la participación del sistema IR-n en las conferencias TREC e iCLEF dentro de las tareas citadas.

En el siguiente capítulo se presentarán las conclusiones finales y las líneas de investigación que se han abierto a partir del trabajo realizado en esta tesis.

7. Conclusiones finales

Es curioso comprobar el increíble cambio que se ha producido en cuanto a la disponibilidad de información en formato digital en los últimos años. Muchos son los factores que han provocado este cambio, no obstante el que mayor influencia ha tenido, sin duda alguna, ha sido Internet. En pocos años se ha pasado de utilizar archivos de fichas de papel para localizar un libro de entre una colección de miles, a poder realizar búsquedas entre miles de millones de documentos disponibles en la red.

Este incremento de la cantidad de información disponible en formato digital ha supuesto un incremento notable del interés en la investigación en sistemas de información textual. El objetivo es claro, se desea automatizar el proceso de búsqueda de todo tipo de información en una serie de documentos. Este proceso de búsqueda ha evolucionado desde intentar localizar documentos que contienen información relevante (sistemas de recuperación de información, RI) a los ambiciosos proyectos de localizar y generar información estructurada (sistemas de extracción de información, EI) o incluso a localizar respuestas concretas en grandes colecciones de documentos (sistemas de búsqueda de respuestas, BR).

Es evidente que los posibles logros de estos dos últimos tipos de sistemas son mucho más ambiciosos que los alcanzados por los sistemas de RI. No obstante, no hay que olvidar un aspecto muy importante. Tanto los sistemas de EI como los sistemas de BR requieren o utilizan sistemas de RI para filtrar los documentos que no contienen información relevante.

Así, el campo de investigación en RI sigue abierto, no sólo en cuanto a mejora de las técnicas para determinar con mayor precisión que documentos son relevantes, sino también con el ob-

jetivo de detectar aquellas partes de un documento que realmente son relevantes para una pregunta o tema determinado. Esta es la línea que siguen los sistemas de RI basados en pasajes (RP). Estos modelos se basan en el estudio de la relevancia de pequeños fragmentos de texto, denominados pasajes, para determinar la relevancia del documento ante una pregunta. Los sistemas de RP, además de aportar una serie de ventajas en los cálculos de relevancia, permiten localizar aquellos fragmentos relevantes de texto que contiene un documento.

El trabajo de investigación desarrollado en esta tesis profundiza en el estudio de los sistemas de RP y valora las ventajas que aportan sobre los sistemas de RI que se basan en el estudio del documento completo (RID). En concreto, se ha definido un nuevo modelo de RP (IR-n) que utiliza la frase como unidad para definir los pasajes. Dadas las características del modelo desarrollado, se ha comprobado los buenos resultados que permite obtener tanto en tareas de RI, como de BR y Selección Interactiva de Documentos (SID).

7.1 Aportaciones

A continuación se resumen las principales contribuciones del trabajo desarrollado en esta tesis:

- Recopilación y estudio de los sistemas de RI basados en el análisis del documento completo (RID).

  Se han descrito las características básicas que debe tener un sistema de RI. Se han definido todos los procesos que deben llevar a cabo. También se ha realizado un estudio de las principales aproximaciones utilizadas en los modelos de cálculo de relevancia.

- Recopilación, estudio y clasificación de los sistemas de RP.

  Los modelos de RP son una alternativa a los modelos RID. En esta tesis se ha realizado un profundo estudio de estos sistemas, indicando sus ventajas e inconvenientes sobre los modelos RID. Las principales ventajas que aportan los sistemas de RP son los siguientes:

  1. Permiten valorar la proximidad de aparición de los términos de la pregunta en el cálculo de similitud.
  2. Detectan no sólo si un documento es relevante, sino que además permiten indicar qué parte del documento realmente lo es.

  Finalmente se ha realizado una clasificación y comparación de las principales propuestas de sistemas de RP.

- Definición de un nuevo modelo de RP

  El sistema IR-n, es un modelo de RP, que se diferencia claramente del resto, principalmente, por la unidad que utiliza para definir los pasajes en los que se divide el documento: la frase.

  Se han analizado las ventajas que supone el uso de la frase como unidad de definición de los pasajes, dentro de la problemática de la RI. Las principales son:

  1. Los límites que definen una frase dentro de un documento se pueden obtener fácilmente aunque no se disponga de marcas en el documento que las identifiquen.
  2. Los pasajes generados en base a un número de frases, están dotados de entidad sintáctica. Esta característica ha sido de gran utilidad en el uso del sistema como paso previo a la aplicación de un sistema de BR, y también, en tareas de SID.
  3. Permiten incorporar de forma sencilla el concepto de solapamiento en la definición de los pasajes.
  4. Permite valorar, en el cálculo de relevancia, el hecho de que los términos que forman la pregunta aparezcan en una misma

unidad como es la frase.

- Estudio y propuesta de nuevas medidas de similitud adecuadas al sistema.

  Se ha efectuado un estudio de las principales medidas de similitud que permiten obtener la relevancia de un documento con respecto a una pregunta. En primer lugar se ha adaptado alguna de éstas (medida del coseno) al sistema IR-n (ver subsección 4.2.2). Posteriormente se ha definido una nueva medida de similitud (ver subsección 4.4.1) más adecuada al sistema. Esta medida es de fácil implementación y permite valorar el hecho de que los términos de la pregunta se hallen en una misma unidad sintáctica y semántica como es la frase. Además, se ha comprobado que esta nueva medida supone un incremento notable en la eficacia del sistema sin provocar un incremento sensible de la complejidad del mismo.

- Definición de un modelo para el tratamiento de preguntas largas en un sistema de RP.

  Las mejoras que el sistema IR-n aporta frente a otros modelos de RI eran menores si las preguntas a procesar estuviesen formadas por más de una frase (preguntas largas). Por ello, se realizó un estudio exhaustivo de este problema y se definió un nuevo modelo de tratamiento de las preguntas largas (ver subsección 4.4.3), con el cual se mejoran considerablemente los resultados obtenidos.

- Estudio y aplicación de técnicas de expansión de la pregunta.

  Se ha analizado las principales técnicas de expansión de la pregunta, tanto la descripción general del modelo que proponen, como las ventajas e inconvenientes de cada uno de ellos. A partir de este análisis se ha estudiado la forma más adecuada de incorporar estas técnicas al sistema IR-n (ver subsección 4.4.2).

- Evaluación de los sistema de RI más importantes.

  Se ha evaluado el sistema IR-n en el ámbito de la conferencia CLEF-2002, lo que ha permitido contrastar los resultados frente a los sistemas de RI más importantes. Esta evaluación ha permitido demostrar la capacidad del sistema IR-n en tareas de RI. Ha sido el quinto mejor sistema a nivel de la media de precisión, y el mejor sistema de todos los participantes a nivel de precisión obtenida a los cinco documentos recuperados.

- Evaluación de los sistema de RP más conocidos.

  Se ha evaluado el sistema IR-n frente a los modelos más conocidos de RP. Esta evaluación muestra los buenos resultados que obtiene el sistema IR-n respecto al resto de modelos de RP, tanto a nivel de media de precisión como a nivel de precisión a los cinco documentos recuperados (en la que es el sistema que mejores resultados obtiene). Además cabe destacar la ventaja que tiene el sistema IR-n frente a las otras propuestas que obtienen resultados similares, y es su uso de pasajes con entidad sintáctica.

- Evaluación del sistema IR-n en tareas de BR.

  El sistema IR-n no es un sistema de BR, pero permite localizar los pasajes más relevantes para una pregunta en una colección. Esto facilita un posterior tratamiento para la realización del proceso de BR.
  En primer lugar, se ha realizado un estudio para determinar de qué forma puede ser más eficaz el sistema IR-n dentro de esta tarea. Posteriormente, se ha comparado frente a un modelo estándar de RI en la tarea de seleccionar textos que contienen las respuestas a una serie de preguntas. En esta comparativa se ha verificado que el sistema IR-n es capaz de reducir considerablemente la cantidad de información que debe procesar un sistema de BR, permitiendo incrementar la eficacia del mismo.

Finalmente, también se ha descrito los resultados obtenidos en la participación en la conferencia TREC-2001, del sistema IR-n junto a un sistema de BR (SEMQA).

- Evaluación del sistema IR-n en tareas de SID.

  Dentro de las tareas de SID se pretende estudiar la forma en la que se puede facilitar a un usuario la determinación de la relevancia o no de un documento frente a una pregunta. El sistema IR-n es capaz de determinar los pasajes más relevantes dentro de una colección de documentos. Dada esta filosofía, se ha estudiado si mostrar estos pasajes más relevantes es suficiente para que un usuario determine la relevancia del documento que los contiene, con lo cual se evitaría el estudio del documento completo. Este estudio se ha realizado en la conferencia iCLEF-2002. También se han contrastado los resultados del sistema IR-n (mostrando sólo los pasajes más relevantes) y un sistema de RI estándar (mostrando el documento completo). Los resultados permitieron comprobar los beneficios que aportaba el sistema IR-n frente al segundo.

7.2 Trabajos en progreso

El objetivo desde el inicio del desarrollo del sistema IR-n, siempre ha sido el incremento de sus prestaciones. Fundamentalmente el incremento de la eficacia y flexibilidad del sistema, sin que esto suponga un menoscabo respecto a su eficiencia.

Hay una serie de elementos que pueden incrementar esa eficacia y que están siendo objeto de estudio actualmente. Las líneas de estudio en progreso son las siguientes:

- Estudio de técnicas de expansión de la pregunta.

  Las técnicas de expansión de la pregunta aplicadas al sistema IR-n han sido las basados en *Thesaurus* y las de *Realimentación*. Estas técnicas han permitido mejorar los resultados del

sistema IR-n. No obstante, esta mejoría dista aún de la conseguida por otros modelos. Por ello, se considera conveniente continuar con el estudio de estas técnicas y su aplicación al sistema IR-n. Actualmente se plantea la posibilidad de incorporar técnicas de desambiguación del sentido de las palabras (Word Sense Disambiguation - WSD) en el primero de los modelos estudiados. Otra línea de investigación que se abre es la idea de combinar las dos aproximaciones con el objeto de completar la pregunta original del usuario para realizar una recuperación de documentos empleando las ventajas que cada una de ellas reporta.

- Integración de técnicas de etiquetado de entidades.

  La integración de este tipo de herramientas permitiría determinar las relaciones existentes entre las palabras que forman la pregunta, y por lo tanto, valorar la aparición de las entidades encontradas en un intento de incrementar la precisión del sistema.

- Determinación de límites de relevancia.

  En los trabajos realizados en la tarea de BR, el sistema IR-n suministra un número fijo de pasajes para cada pregunta. Esto puede provocar que para algunas preguntas se incorporen pasajes con baja relevancia, o que en otros casos, no se consideren pasajes de similar relevancia a los últimos suministrados. Se considera interesante la posibilidad de realizar un estudio para cada pregunta en función de la valoración obtenida de cada pasaje, con el objeto de adecuar el número de pasajes devueltos para cada pregunta.

- Estudio del uso de pasajes de tamaño variable.

  Otro objetivo es el de dotar al sistema de la posibilidad de utilizar pasajes de tamaño variable. Esto obligaría a replantear el método de cálculo de relevancia, dotándolo de una normali-

zación que facilite comparar pasajes de distinto tamaño. Este modelo también permitiría utilizar pasajes de distinto tamaño en función de la colección utilizada en cada momento. Esto es de utilidad en colecciones como las del TREC, que están formadas por varias colecciones de documentos de diferentes características.

7.3 Publicaciones realizadas

Los resultados presentados en esta tesis han sido contrastados en foros de investigación como publicaciones en revistas y congresos internacionales. A continuación se presentan los trabajos publicados durante el transcurso de esta tesis.

El resumen de publicaciones es: 4 artículos en revistas nacionales, 10 artículos en congresos internacionales y 1 informe técnico. Los artículos se presentan según el tipo de publicación.

(i) Artículos en revistas nacionales.

- F. Llopis, A. Ferrández, J. L. Vicedo. "Selección de pasajes para facilitar el proceso de búsqueda de respuestas". Procesamiento Lenguaje Natural. Número 29, Pags. 273-280 , Septiembre 2002

- F. Llopis, A. Ferrández, J. L. Vicedo. "Utilización de pasajes de tamaño variable para mejorar el proceso de recuperación de información". Procesamiento Lenguaje Natural. Número 28, Pags. 89-98, Mayo 2002

- R. Muñoz, A. Montoyo, F. Llopis, A. Suárez. "Reconocimiento de Entidades en el sistema EXIT". Procesamiento Lenguaje Natural. Número 23, Pags. 47-53, Septiembre 1998

- F. Llopis, R. Muñoz, A. Suárez, A. Montoyo. "EXIT: Propuesta de un sistema de extracción de información de textos

notariales". Revista NOVÁTICA. Número 133, Pags. 26-30, Mayo-Junio 1998.

(ii) Artículos en congresos internacionales.

- J. L. Vicedo, F. Llopis, A. Ferrández "University of Alicante. Experiments at TREC-2002". Eleventh Text REtrieval Conference **(TREC-11)**. Gaithersburg, Maryland (EEUU). November 2002 . NIST Special Publication 500-250. Gaithersburg, US. Noviembre 2002.

- F. Llopis, J. L. Vicedo, A. Ferrández. "IR-n System at Clef-2002". Workshop of The Cross-Language Evaluation Forum (**Clef 2002**)". Roma, Italy. Lecture Notes in Computer Science. Springer-Verlag, Pendiente de publicación

- F. Llopis, A. Ferrández, J. L. Vicedo, M. Díaz, F. Martínez. "iCLEF at Universities of Alicante and Jaen". Workshop of The Cross-Language Evaluation Forum (**Clef 2002**)". Roma, Italy. Lecture Notes in Computer Science. Springer-Verlag, Pendiente de publicación

- F. Llopis, A. Ferrández, J. L. Vicedo. "Passage election to Improve Question Answering" Proceedings of the Workshop on Multilingual Summarization and Question Answering. **(COLING 2002)** Post-Conference Workshops. Taipei, Taiwan. Pags. 11-16

- F. Llopis, A. Ferrández, J. L. Vicedo. "Using a Passage Retrieval System to Support Question Answering Process". The 2002 International Conference on Computational Science (**ICCS 2002**)". Amsterdam, The Netherlands. Lecture Notes in Computer Science. Springer-Verlag, Abril 2002. Volumen 2329. Pags. 61-69

- F. Llopis, J. L. Vicedo, A. Ferrández. "Using long queries in a passage retrieval system". Mexican International

Conference on Artificial Intelligence (**MICAI 2002**)". Mexico City, Mexico. Lecture Notes in Artificial Intelligence. Springer-Verlag, Abril 2002. Volumen 2313. Pags. 185-193.

- F. Llopis, J. L. Vicedo, A. Ferrández. "Text Segmentation for efficient Information Retrieval". Third International Conference on Intelligent Text Processing and Computational Linguistics (**CICLing 2002**)". Mexico City, Mexico. Lecture Notes in Computer Science. Springer-Verlag, Febrero 2002. Volumen 2276. Pags. 373-380.

- J. L. Vicedo, A. Ferrández, F. Llopis. "University of Alicante at TREC-10". Tenth Text REtrieval Conference (**TREC-10**). NIST Special Publication 500-250. Gaithersburg, US. Noviembre 2001.

- F. Llopis, J. L. Vicedo. "IR-n System: A Passage Retrieval System at Clef-2001". Workshop of The Cross-Language Evaluation Forum (**Clef 2001**)". Darmstadt, Germany. Lecture Notes in Computer Science. Springer-Verlag, Septiembre 2001. Volumen 2406. Pags. 244-252.

- J. L. Vicedo, F. Llopis, A. Ferrández. "De la recuperación de información a los sistemas de búsqueda de respuestas o Question Answering". Segundo Taller Internacional de Procesamiento Computacional del Español y Tecnologías del Lenguaje (**SLPLT2**). Pags. 89-94. Jaén, España. Septiembre 2001.

(iii) Informe Técnico

- F. Llopis, R. Muñoz, A. Suárez, A. Montoyo, M. Palomar, A. Ferrández, P. Martínez-Barco, J. Peral, R. Romero, M. Saiz. "Propuesta de un sistema de extracción de información de textos notariales". Report Interno. Departamento de Lenguajes y Sistemas Informáticos.

Bibliografía

ALFONSECA, E., M. DE BONI, J.L. JARA-VALENCIA y S. MANANDHAR (2002). «A prototype Question Answering System using syntactic and semantic information for answer retrieval», en TREC-10 (2002), págs. 680–685.

ALPHA, S., P. DIXON, C. LIAO y C. YANG (2002). «Oracle at TREC 10», en TREC-10 (2002), págs. 423–433.

ATTARDI, G., A. CISTERNINO, F. FORMICA, M. SIMI, A. TOMMASI y C. ZAVATTARI (2002). «PIQASso: PIsa Question Answering System», en TREC-10 (2002), págs. 633–641.

BAEZA-YATES, RICARDO y BERTHIER RIBEIRO-NETO (1999). *Modern Information Retrieval*, cap. 10: User Interfaces and Visualization, págs. 257–323, Addison Wesley, Reading, US.

BRAND, ROEL y MARVIN BRÜNNER (2002). «Océ at Clef-2002», en CLEF (2002), págs. 21–30.

BRILL, E., J. LIN, M. BANKO y S. DUMAIS (2002). «Data-Intensive Question Answering», en TREC-10 (2002), págs. 393–400.

BROWN, G. y G. YULE (1983). *Discourse Analisys*, Cambridge University Press, London New York.

BUCHHOLZ, SABINE (2002). «Using Grammatical Relations, Answer Frequencies and the World Wide Web for TREC Question Answering», en TREC-10 (2002), págs. 502–509.

BUCKLEY, C. (1994). «The importance of proper weighting methods», en *Proc. ARPA Human Language Technology Workshop '93*, págs. 349–352, Princeton, NJ, distributed as *Human Language Technology* by San Mateo, CA: Morgan Kaufmann Publishers.

BUCKLEY, C., A. SINGHAL, M. MITRA y G. SALTON (1995). «New retrieval approaches using smart: Trec», en *Fourth Text REtrieval Conference*, vol. 500-238 de *NIST Special Publication*, págs. 25–48, National Institute of Standards and Technology, Gaithersburg, USA.

BUCKLEY, CHRIS, GERARD SALTON, JAMES ALLAN y AMIT SINGHAL (1994). «Automatic query expansion using SMART: TREC 3», en *Text REtrieval Conference*, págs. 69–80.

C. FIGUEROLA, A. ZAZO, R. GÓMEZ y J.L. BERROCAL (2001). «Stemming in Spanish: A First Approach to its Impact on Information Retrieval», en CLEF (2001), págs. 253–261.

CALLAN, JAMES P. (1994). «Passage-Level Evidence in Document Retrieval», en *Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval*, págs. 302–310, Springer Verlag, London, UK.

CALLAN, JAMES P., W. BRUCE CROFT y STEPHEN M. HARDING (1992). «The INQUERY retrieval system», en *Proceedings of DEXA-92, 3rd International Conference on Database and Expert Systems Applications*, págs. 78–83.

CASSANY, DANIEL (1990). «Enfoques didácticos para la enseñanza de la expresión escrita», *Comunicación, lenguaje y educación*, **6**, 63–80.

CHEN, AITAO (2002). «Cross-Languaje Retrieval Experiments at CLEF-2002», en CLEF (2002), págs. 5–20.

CHEN, J., A. DIEKEMA, M. TAFFET, N. MCCRACKEN, N. OZGENCIL, O. YILMAZEL y E. LIDDY (2002). «Question Answering: CNLP at the TREC-10 Question Answering Track», en TREC-10 (2002), págs. 485–494.

CLARKE, CHARLES L., G. V. CORMACK, T. R. LYNAM, C. M. LI y G. L. MCLEAN (2002). «Web Reinforced Question Answering (MultiText Experiments for TREC 2001)», en TREC-10 (2002), págs. 673–669.

CLEF (2001). «Workshop of cross-language evaluation forum (clef 2001)», en *Workshop of Cross-Language Evaluation Forum (CLEF 2001)*, Lecture notes in Computer Science, Springer-Verlag, Darmstadt, Germany.

CLEF (2002). *Workshop of Cross-Language Evaluation Forum (CLEF 2002)*, Lecture notes in Computer Science, Roma, Italy, Springer-Verlag.

COOPER, WILLIAM S. (1991). «Some inconsistencies and misnomers in probabilistic information retrieval», en *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, págs. 57–61, ACM Press.

COWIE, J. R. y W. LEHNERT (1996). «Information Extraction», *Comunications of ACM*, **39**(1), 81–91.

CROFT, W. BRUCE (1995). «What do people want from information retrieval? (the top 10 research issues for companies that use and sell IR systems)», *D-Lib Magazine*.

DEITEL, H. C. y P. J. DEITEL (1999). *C++ Como programar.*, Prentice-Hall, México.

FERRET, O., B. GRAU, M. HURAULT-PLANTET, G. ILLOUZ, L. MONCEAUX y A. VILNAT (2002). «Finding an answer based on the recognition of the question focus», en TREC-10 (2002), págs. 362–370.

FOX, C. (1992). *Lexical Analysis and Stoplists*, cap. 7, págs. 102–130, en Frakes y Baeza-Yates (1992).

FRAKES, W. B. y R. BAEZA-YATES (1992). *Information Retrieval: Data Structures & Algorithms*, Prentice Hall, Englewood Cliffs, N.J.

GONZALO, JULIO y DOUGLAS W. OARD (2002). «The CLEF 2002 Interactive Track», en CLEF (2002), págs. 245–253.

HARABAGIU, SANDA, DAN MOLDOVAN, MARIUS PASCA, RADA MIHALCEA, MIHAI SURDEANU, RAZVAN BUNESCU, ROXANA GÎRJU, VASILE RUS, PAUL MORARESCU y FINLEY LACATUSU (2002). «Answering complex, list and context questions with LCC's Question-Answering Server», en TREC-10 (2002), págs. 355–361.

HARMAN, DONNA (1988). «Towards Interactive Query Expansion», en *Proceedings of the Eleventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Applications (2), págs. 321–331, Grenoble, France.

HARMAN, DONNA (1992a). *Relevance feedback and other query modification techniques*, págs. 241–263, en Frakes y Baeza-Yates (1992).

HARMAN, DONNA (1992b). «Relevance feedback revisited», en *Fifteenth International ACM SIGIR Conference on Research and Development in Information Retrieval*, Interaction in Information Retrieval, págs. 1–10, New York, USA.

HEARST, M. y C. PLAUNT (1993). «Subtopic structuring for full-length document access», en SIGIR (1993), págs. 59–68.

HEARST, MARTI (1994). «Multi-paragraph segmentation of expository text», en *32nd. Annual Meeting of the Association for Computational Linguistics*, págs. 9–16, New Mexico State University, Las Cruces, New Mexico.

HENSTOCK, PETER V., DANIEL J. PACK, YOUNG-SUK LEE y CLIFFORD J. WEINSTEIN (2001). «Toward an improved concept-based information retrieval system», en *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, págs. 384–385, ACM Press.

HOVY, E., U. HERMJACOB y C. LIN (2002). «The Use of External Knowledge in Factoid QA», en TREC-10 (2002), págs. 644–652.

ITTYCHERIAH, ABRAHAM, MARTIN FRANZ y SALIM ROUKOS (2002). «IBM's Statistical Question Answering System - TREC-10», en TREC-10 (2002), págs. 258–264.

J. BROGLIO, B. CROFT, J. CALLAN y D.ÑACHBAR (1994). «Document retrieval and routing using the inquery system», **500-238**, 29–38.

JONES, KAREN SPARCK (1981). «Retrieval system tests 1958 — 1978», en Karen Sparck Jones, editor, *Information Retrieval Experiment*, págs. 213–255, pub-BUTTERWORTHS.

JOURLIN, P., S.E. JOHNSON, K. SPÄRCK JONES y P.C. WOODLAND (1999). «General query expansion techniques for spoken document retrieval», en *Proc. ESCA Workshop on Extracting Information from Spoken Audio*, págs. 8–13, Cambridge, UK.

K. RICHMOND, A. SMITH y E. AMITAY (1997). «Detecting Subject Boundaries Within Text: A Language Independent Statisti-

cal Approach», en *In proceedings of The Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*, págs. 1–8, University of Brown, Rhode Island, USA.

KASZKIEL, MARCIN y JUSTIN ZOBEL (1997). «Passage Retrieval Revisited», en *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Text Structures, págs. 178–185, Philadelphia, PA, USA.

KASZKIEL, MARCIN y JUSTIN ZOBEL (2001). «Effective Ranking with Arbitrary Passages», *Journal of the American Society for Information Science (JASIS)*, **52**(4), 344–364.

KASZKIEL, MARCIN, JUSTIN ZOBEL y RON SACKS-DAVIS (1999). «Efficient passage ranking for document databases», *ACM Transactions on Information Systems*, **17**(4), 406–439.

KAZAWA, H., H. ISOZAKI y E. MAEDA (2002). «NTT Question Answering System in TREC 2001», en TREC-10 (2002), págs. 415–422.

KWOK, K., L. GRUNFELD, N. DINSTL y M. CHAN (2002). «TREC 2001 Question-Answer, Web and Cross Language Experiments using PIRCS», en TREC-10 (2002), págs. 452–456.

LAM-ALESINA, ADENIKE y GARETH JONES (2002). «Exeter at Clef-2002. Experiments with Machine Translation for Monolingual and Bilingual Retrieval», en CLEF (2002), págs. 63–72.

LANCASTER, F. (1968). *Information Retrieval Sytems. Characteristics, Testing and Evaluation*, Wiley NewYork.

LEE, G. GEUNBAE, S. LEE, H. JUNG, B-H CHO, C. LEE, B-K KWAK, J. CHA, D. KIM, J. AN, J. SEO, H. KIM y K. KIM (2002). «SiteQ: Engineering High Performance QA System Using Lexico-Semantic Pattern Matching and Shallow NLP», en TREC-10 (2002).

LIN, CHUAN-HIE y HSIN-HSI CHEN (2002). «Description of NTU System at TREC-10 QA Track», en TREC-10 (2002).

LITKOWSKI, K.C. (2002). «CL Research Experiments in TREC-10 Question Answering», en TREC-10 (2002), págs. 122–131.

LLOPIS, F., A. FERRÁNDEZ y J.L. VICEDO (2002a). «Selección de pasajes para facilitar el proceso de búsqueda de respuestas», *Procesamiento del Lenguaje Natural*, **29**, 273–280.

LLOPIS, F., A. FERRÁNDEZ y J.L. VICEDO (2002b). «Utilización de pasajes de tamaño variable para mejorar el proceso de recuperación de información», *Procesamiento del Lenguaje Natural*, **28**, 89–98.

LLOPIS, F., R. MUÑOZ, A. SUÁREZ y A. MONTOYO (1998). «EXIT: Propuesta de un sistema de extracción de información de textos notariales», *Novática*, **133**, 26–30.

LLOPIS, FERNANDO, , ANTONIO FERRÁNDEZ y JOSÉ L. VICEDO (2002c). «Passage Selection to Improve Question Answering», en ACL, editor, *Proceedings of the 19th Annual Conference on Computational Linguistics*, págs. 11–16, Taipei, Taiwan.

LLOPIS, FERNANDO, ANTONIO FERRÁNDEZ y JOSÉ L. VICEDO (2002d). «Using a Passage Retrieval System to Support Question Answering Process», en *The 2002 International Conference on Computational Science (ICCS 2002)*, Lecture notes in Computer Science, Springer-Verlag, Amsterdam, The Netherlands.

LLOPIS, FERNANDO, ANTONIO FERRÁNDEZ y JOSÉ L. VICEDO (2002e). «Using Long Queries in a Passage Retrieval System», en O. Cairo, E. L. Sucar y F. J. Cantu, editores, *Proceeding of Mexican International Conference on Artificial Intelligence*, vol. 2313 de *Lectures Notes in Artificial Intelligence*, Springer-Verlag, Mérida, Mexico.

LLOPIS, FERNANDO y JOSÉ L. VICEDO (2001). «IR-n system, a passage retrieval system at CLEF 2001», en CLEF (2001), págs. 244–252.

LLOPIS, FERNANDO, JOSÉ L. VICEDO y ANTONIO FERRÁNDEZ (2002f). «Text Segmentation for efficient Information Retrieval», en *Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002)*, Lecture notes in Computer Science, págs. 373–380, Springer-Verlag, Mexico City, Mexico.

LLOPIS, FERNANDO, JOSÉ LUIS VICEDO y ANTONIO FERRÁNDEZ (2002g). «IR-n system at Clef-2002», en CLEF (2002), págs. 177–184.

LLOPIS, FERNANDO, JOSÉ LUIS VICEDO, ANTONIO FERRÁNDEZ, MANUEL DÍAZ y FERNANDO MARTÍNEZ (2002h). «iClef at Universities of Alicante and Jaen», en CLEF (2002), págs. 261–266.

MAGNINI, B., M.ÑEGRI, R. PREVETE y H. TANEV (2002). «Multilingual Question/Answering: the DIOGENE System», en TREC-10 (2002), págs. 313–321.

MCFARLANE, ANDREW (2002). «PLIERS and SnowBall at CLEF-2002», en CLEF (2002), págs. 191–198.

MCNAMEE, PAUL y JAMES MAYFIELD (2002). «Scalable Multilingual Information Access», en CLEF (2002), págs. 133–140.

MILLER, G. (1995). «Wordnet: A Lexical Database for English», en *Communications of the ACM 38(11))*, págs. 39–41.

MILLER, G., R. BECKWITH, C. FELLBAUM, D. GROSS y K. MILLER (1990). «Five Papers on WordNet», *CLS Rep. 43*, Princeton University, Cognitive Science Laboratory.

MOFFAT, ALISTAIR, RON SACKS-DAVIS, ROSS WILKINSON y JUSTIN ZOBEL (1993). «Retrieval of partial documents», en *Text REtrieval Conference*, págs. 181–190.

MOFFAT, ALISTAIR y JUSTIN ZOBEL (1996). «Self-indexing inverted files for fast text retrieval», *ACM Transactions on Information Systems*, **14**(4), 349–379.

MONZ, CHRISTOF y MAARTEN DE RIJKE (2002). «Tequesta: The University of Amsterdam's Textual Question Answering System», en TREC-10 (2002), págs. 519–528.

MONZ, CHRISTOF, JAAP KAMPS y MAARTEN DE RIJKE (2002). «The University of Amsterdam at Clef-2002», en CLEF (2002), págs. 73–84.

MOULINIER, ISABELLE y HUGO MOLINA-SALGADO (2002). «Thomson Legal and Regulatory Experiments for CLEF 2002», en CLEF (2002), págs. 91–96.

MUÑOZ, R., A. MONTOYO, F. LLOPIS y A. SUÁREZ (1998). «Reconocimiento de entidades en el sistema EXIT», *Procesamiento del Lenguaje Natural*, **23**, 47–53.

MUÑOZ, R. y M. PALOMAR (1999). *Emerging Technologies in Accounting and Finance*, cap. Sentence Boundary and Named

Entity Recognition in EXIT System: Information Extraction System of Notarial Texts, págs. 129–142.

OH, J., K. LEE, D. CHANG, C. WON SEO y K. CHOI (2002). «TREC-10 Experiments at KAIST: Batch Filtering and Question Answering», en TREC-10 (2002), págs. 347–354.

PLAMONDON, L., G. LAPALME y L. KOSSEIM (2002). «The QUANTUM Question Answering System», en TREC-10 (2002), págs. 579–585.

PONTE, J. M. y W. B. CROFT (1997). «Text segmentation by topic», *Lecture Notes in Computer Science*, **1324**, 113–125.

PORTER, M. (1980). «An algorithm for suffix stripping», *Program-automated library and information systems*, **14**(3), 130–137.

PRAGER, JOHN, JENNIFER CHU-CARROLL y KRZYSZTOF CZUBA (2002). «Use of Wordnet Hypernyms for Answering What-Is Questions», en TREC-10 (2002), págs. 250–257.

QIU, YONGGANG y H.P. FREI (1993). «Concept based query expansion», en SIGIR (1993), págs. 160–169.

RENNERT, P. (2002). «TREC 2001 - Word Proximity QA System», en TREC-10 (2002).

RIJSBERGEN, C. J. VAN (1979). *Information Retrieval, 2nd edition*, Butterworths, London.

ROBERSTON, S.E. (1977). «The probability ranking principle in ir», *Journal of Documentation*, **33**(4), 294–304.

ROBERSTON, S.E., S. WALKER y M. BEAULIEU (2000). «Okapi at trec», *Information Processing and Management*, **36**(1), 95–108.

ROBERTSON, S. E. y K. SPARCK JONES (1976). «Relevance weighting of search terms.», *J. Amer. Soc. for Information Sci.*, **27**, 129–146.

ROTH, D., G. KAO, X. LI, R.ÑAGARAJAN, V. PUNYAKANOK, N. RIZZOLO, W. YIH, C OVESDOTTER y L. GERARD (2002). «Learning Components for a Question-Answering System», en TREC-10 (2002), págs. 539–548.

S. ROBERSTON, S. WALKER y M. BEAULIEU. (1998). «okapi at TREC-7», en TREC-7 (1998), págs. 253–264.

SAGAN, CARL (1982). *Cosmos*, Planeta, Madrid.

SALTON, G. y J. ALLAN (1994). «Automatic Text Decomposition and Structuring», en SIGIR (1993), págs. 49–58.

SALTON, G., J. ALLAN y C. BUCKLEY (1993). «Approaches to passage retrieval in full text information systems», en SIGIR (1993), págs. 49–58.

SALTON, G., E. FOX y H. WU (1983). «Extnded boolean information retrieval», *Communications of ACM*, **11**, 1022–1036.

SALTON, GERARD y CHRIS BUCKLEY (1988). «A term-weighting approaches in automatic text retrieval», *Information Processing and Management*, **24**(5), 513–123.

SALTON, GERARD, AMIT SINGHAL, CHRIS BUCKLEY y MANDAR MITRA (1996). «Automatic text decomposition using text segments and text themes», en *Proceedings of the Seventh ACM Conference on Hypertext*, Autonomous Hypertext Systems and Link Discovery, págs. 53–65.

SALTON, GERARD A. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison Wesley, New York.

SALTON, GERARD A. y M. J. MCGILL (1983). *Introduction to Modern Information Retrieval*, McGraw-Hill, Tokio.

SAVOY, JACQUES (2001). «Report on CLEF-2001 Experiments:Effective Combined», en CLEF (2001), págs. 27–43.

SAVOY, JACQUES (2002). «Report on CLEF-2002 », en CLEF (2002), págs. 27–43.

SIGIR (1993). *Sixteenth International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, PA.

SINGHAL, AMIT, CHRIS BUCKLEY y MANDAR MITRA (1996). «Pivoted document length normalization», en *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Experimental Studies, págs. 21–29.

SINGHAL, AMIT, JOHN CHOI, DONALD HINDLE, DAVID D. LEWIS y FERNANDO C.Ñ. PEREIRA (1998). «ATT at TREC-7», en *Text REtrieval Conference*, págs. 186–198.

SMEATON, ALAN F. (1997). «Information retrieval: Still butting heads with natural language processing?», en *SCIE*, págs. 115–138.

SOUBBOTIN, M. y S. SOUBBOTIN (2002). «Patterns of Potential Answer Expressions as Clues to the Right Answers», en TREC-10 (2002), págs. 293–302.

SPARCK-JONES, KAREN (1972). «A statistical interpretation of term specificity and its application in retrieval», *Journal of Documentation*, **28**, 11–21.

SPARK-JONES, KAREN (1999). «What is the Role of NLP in Text Retrieval?», en *Natural Language Information Retrieval*, cap. 1, págs. 1–24, Kluwer Academic, New York, USA.

STRZALKOWSKI, T., G. STEIN, G. BOWDEN WISE, J. PEREZ-CARBALLO, P. TAPANANINEN, T. JARVINEN, A. VOUTILAINEN y J. KARLGREN (1998). «Natural language information retrieval: TREC-7 report», en TREC-7 (1998), págs. 217–226.

THOMLINSON, STEPHEN (2002). «Experimentes in 8 European Languajes with Humingbird Search Server», en CLEF (2002), págs. 177–184.

TREC-10 (2002). *Tenth Text REtrieval Conference*, vol. 500-250 de *NIST Special Publication*, Gaithersburg, USA, National Institute of Standards and Technology.

TREC-7 (1998). *Seventh Text REtrieval Conference*, vol. 500-242 de *NIST Special Publication*, Gaithersburg, USA, National Institute of Standards and Technology.

TREC-8 (1999). *Eighth Text REtrieval Conference*, vol. 500-246 de *NIST Special Publication*, Gaithersburg, USA, National Institute of Standards and Technology.

VICEDO, J. (2002). *SEMQA: Un modelo semántico aplicado a los sistemas de Búsqueda de Respuestas*, Tesís doctoral, Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante.

VICEDO, JOSÉ LUIS, ANTONIO FERRÁNDEZ y FERNANDO LLOPIS (2002). «University of Alicante at TREC-10», en TREC-10 (2002), págs. 510–518.

VICEDO, JOSÉ LUIS, FERNANDO LLOPIS y ANTONIO FERRÁNDEZ (2003). «University of Alicante at TREC-2002»,

en *Eleventh Text REtrieval Conference (Notebook)*, vol. 500-250 de *NIST Special Publication*, National Institute of Standards and Technology, Gaithersburg, USA.

VILARES, JESÚS, MIGUEL A. ALONSO, FRANCISCO J.RIBADAS y MANUEL VILARES (2002). «COLE experiments at Clef-2002. Spanish monolingual track», en CLEF (2002), págs. 153–160.

VOORHEES, ELLEN M. (1994). «Query expansion using lexical-semantic relations», en *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Indexing, págs. 61–69.

VOORHEES, ELLEN M. y DONNA HARMAN (1999). «Overview of the Eighth Text REtrieval Conference», en TREC-8 (1999), págs. 1–24.

WANG, B., H. ZU, Z. YANG, Y. LIU, X. CHENG, D. BU y S. BAI (2002). «TREC 10 Experiments at CAS-ICT: Filtering, Web and QA», en TREC-10 (2002), págs. 109–121.

WILKINSON, ROSS (1994). «Effective retrieval of structured documents», en *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Passage Retrieval, págs. 311–317.

WILKS, Y. (1997). «Information extraction as a core language technology», *Lecture Notes in Computer Science*, **1299**, 1–9.

WITTEN, I., A. MOFFAT y T. BELL (1999). *Managing Gygabytes*, Morgan Kauffman.

WOODS, W. A., S. GREEN, P. MARTIN y A. HOUSTON (2002). «Aggressive Morphology and Lexical Relations for Query Expansion», en TREC-10 (2002), págs. 479–484.

WU, L., X. HUANG, Y. GUO, Y. XIA y Z. FENG (2002). «FDU at TREC-10: Filtering, Q&A, Web and Video Tasks», en TREC-10 (2002), págs. 192–207.

XU, JINXI y W. BRUCE CROFT (1996). «Query expansion using local and global document analysis», en *Research and Development in Information Retrieval*, págs. 4–11.

ZAZO, ANGEL F., CARLOS G. FIGUEROLA, JOSÉ L. BERROCAL y EMILIO RODRÍGUEZ (2002). «Term Expansion using Stemming and Thesauri in Spanish», en CLEF (2002), págs. 177–184.

Volver al índice/Tornar a l'índex

ZOBEL, J., A. MOFFAT, R. WILKINSON y R. SACKS-DAVIS (1995). «Efficient retrieval of partial documents», *Information Processing & Management*, **31**(3), 361–377.

A. Resultados completos de los experimentos realizados

Este apéndice presenta en detalle los resultados de los diferentes experimentos desarrollados en la fase de entrenamiento del sistema, que se han incorporado de forma resumida en el el capítulo 5 para facilitar su lectura.

		Precisión a los N documentos					
Frases	**Cobertura**	**5**	**10**	**20**	**30**	**200**	**AvgP**
5	94,86	0,4809	0,3872	0,2947	0,2546	0,0762	0,4622
10	95,21	0,5106	0,4170	0,3096	0,2624	0,0752	**0,4896**
15	95,21	0,5064	0,4043	0,3106	0,2610	0,0737	0,4667
20	94,98	0,5277	0,4191	0,3021	0,2539	0,0740	0,4751
25	94,04	0,5191	0,3979	0,3032	0,2496	0,0745	0,4652
30	93,69	0,5064	0,4043	0,3021	0,2482	0,0740	0,4563
35	93,22	0,4979	0,4000	0,3011	0,2504	0,0738	0,4493
40	92,99	0,4851	0,3872	0,3032	0,2468	0,0736	0,4387

Tabla A.1. Resultados sistema IR-n con diferentes tamaños de pasajes en intervalos de 5 frases. Colección LATimes con preguntas cortas

		Precisión a los N documentos					
Frases	**Cobertura**	**5**	**10**	**20**	**30**	**200**	**AvgP**
5	95,33	0,4936	0,4128	0,3085	0,2624	0,0761	0,4913
10	95,68	0,5277	0,4277	0,3277	0,2660	0,0756	**0,5007**
15	95,21	0,5021	0,4404	0,3213	0,2631	0,0749	0,4965
20	94,51	0,5149	0,4319	0,3160	0,2574	0,0744	0,4993
25	94,04	0,5191	0,4128	0,3138	0,2617	0,0737	0,4856
30	93,34	0,5021	0,4149	0,3064	0,2546	0,0733	0,4841
35	92,76	0,4979	0,4191	0,3011	0,2511	0,0726	0,4739
40	91,59	0,4979	0,4043	0,3000	0,2447	0,0719	0,4637

Tabla A.2. Resultados sistema IR-n con diferentes tamaños de pasajes en intervalos de 5 frases. Colección LATimes con preguntas largas

		Precisión a los N documentos					
Frases	**Cobertura**	**5**	**10**	**20**	**30**	**200**	**AvgP**
5	94,69	0,6286	0,5592	0,4796	0,4449	0,1914	0,4969
10	94,91	0,6735	0,6020	0,5041	0,4510	0,1939	**0,5052**
15	94,65	0,6898	0,6102	0,5041	0,4429	0,1930	0,4963
20	94,65	0,6857	0,6020	0,5000	0,4442	0,1919	0,4917
25	94,69	0,6857	0,5980	0,4918	0,4429	0,1920	0,4908
30	94,65	0,6816	0,5959	0,4918	0,4422	0,1921	0,4889
35	94,73	0,6816	0,5959	0,4908	0,4415	0,1922	0,4885
40	94,69	0,6816	0,5959	0,4908	0,4408	0,1921	0,4884

Tabla A.3. Resultados sistema IR-n con diferentes tamaños de pasajes en intervalos de 5 frases. Colección EFE con preguntas cortas

		Precisión a los N documentos					
Frases	**Cobertura**	**5**	**10**	**20**	**30**	**200**	**AvgP**
5	95,66	0,6735	0,5837	0,4918	0,4463	0,1939	**0,5085**
10	95,21	0,6612	0,6000	0,5071	0,4497	0,1935	0,5007
15	94,58	0,6612	0,5980	0,5143	0,4524	0,1895	0,4819
20	94,25	0,6531	0,5857	0,5041	0,4463	0,1883	0,4728
25	94,25	0,6531	0,5878	0,5020	0,4456	0,1873	0,4716
30	94,25	0,6531	0,5857	0,5000	0,4449	0,1876	0,4709
35	94,28	0,6490	0,5837	0,4980	0,4442	0,1878	0,4700
40	94,28	0,6490	0,5837	0,4980	0,4435	0,1879	0,4698

Tabla A.4. Resultados sistema IR-n con diferentes tamaños de pasajes en intervalos de 5 frases. Colección EFE con preguntas largas

		Precisión a los N documentos					
Frases	**Cobertura**	**5**	**10**	**20**	**30**	**200**	**AvgP**
5	94,86	0,4809	0,3872	0,2947	0,2546	0,0762	0,4622
6	95,21	0,4894	0,3957	0,3043	0,2574	0,0753	0,4780
7	95,09	0,4936	0,4000	0,3117	0,2617	0,0752	0,4750
8	95,21	0,4936	0,4149	0,3138	0,2589	0,0753	0,4744
9	94,98	0,5191	0,4106	0,3106	0,2582	0,0754	0,4736
10	95,21	0,5106	0,4170	0,3096	0,2624	0,0752	**0,4896**
11	95,21	0,5277	0,4149	0,3149	0,2617	0,0751	0,4871
12	94,98	0,5404	0,4128	0,3106	0,2624	0,0745	0,4866
13	94,86	0,5234	0,4128	0,3085	0,2652	0,0749	0,4892
14	95,09	0,5234	0,4106	0,3064	0,2617	0,0740	0,4703
15	95,21	0,5064	0,4043	0,3106	0,2610	0,0737	0,4667

Tabla A.5. Resultados sistema IR-n con diferentes tamaños de pasajes, subintervalo óptimo. Colección LATimes con preguntas cortas

		Precisión a los N documentos					
Frases	**Cobertura**	**5**	**10**	**20**	**30**	**200**	**AvgP**
5	95,33	0,4936	0,4128	0,3085	0,2624	0,0761	0,4913
6	95,91	0,5234	0,4021	0,3117	0,2660	0,0768	**0,5160**
7	96,03	0,5021	0,4106	0,3160	0,2610	0,0766	0,5012
8	95,68	0,5021	0,4255	0,3223	0,2624	0,0766	0,4984
9	95,56	0,5319	0,4319	0,3255	0,2617	0,0764	0,5040
10	95,68	0,5277	0,4277	0,3277	0,2660	0,0756	0,5007
11	95,33	0,5191	0,4340	0,3309	0,2660	0,0751	0,5035
12	95,09	0,5106	0,4298	0,3298	0,2652	0,0756	0,5009
13	95,44	0,5191	0,4319	0,3266	0,2652	0,0752	0,5056
14	95,33	0,5106	0,4298	0,3202	0,2638	0,0750	0,4956
15	95,21	0,5021	0,4404	0,3213	0,2631	0,0749	0,4965

Tabla A.6. Resultados sistema IR-n con diferentes tamaños de pasajes, subintervalo óptimo. Colección LATimes con preguntas largas

		Precisión a los N documentos					
Frases	**Cobertura**	**5**	**10**	**20**	**30**	**200**	**AvgP**
hline 5	94,69	0,6286	0,5592	0,4796	0,4449	0,1914	0,4969
6	94,73	0,6408	0,5633	0,4939	0,4463	0,1952	0,5081
7	95,03	0,6735	0,5776	0,4969	0,4469	0,1947	0,5071
8	95,10	0,6694	0,5898	0,4990	0,4463	0,1940	**0,5093**
9	94,95	0,6612	0,6041	0,4990	0,4524	0,1936	0,5064
10	94,91	0,6735	0,6020	0,5041	0,4510	0,1939	0,5052
11	94,88	0,6776	0,6000	0,5020	0,4497	0,1938	0,5046
12	94,77	0,6857	0,6082	0,5010	0,4544	0,1931	0,5011
13	94,77	0,6898	0,6143	0,5020	0,4497	0,1932	0,4993
14	94,65	0,6816	0,6082	0,5020	0,4497	0,1941	0,4986
15	94,65	0,6898	0,6102	0,5041	0,4429	0,1930	0,4963

Tabla A.7. Resultados sistema IR-n con diferentes tamaños de pasajes, subintervalo óptimo. Colección EFE con preguntas cortas

		Precisión a los N documentos					
Frases	**Cobertura**	**5**	**10**	**20**	**30**	**200**	**AvgP**
5	95,66	0,6735	0,5837	0,4918	0,4463	0,1939	0,5085
6	96,10	0,6653	0,5857	0,4980	0,4395	0,1970	0,5140
7	96,14	0,6612	0,5959	0,5071	0,4517	0,1967	**0,5178**
8	95,88	0,6776	0,6000	0,5051	0,4497	0,1956	0,5149
9	95,36	0,6612	0,5857	0,5071	0,4592	0,1951	0,5075
10	95,21	0,6612	0,6000	0,5071	0,4497	0,1935	0,5007
11	94,77	0,6571	0,6020	0,5102	0,4544	0,1930	0,4967
12	94,77	0,6490	0,5980	0,5122	0,4571	0,1914	0,4904
13	94,58	0,6653	0,6000	0,5163	0,4551	0,1906	0,4884
14	94,54	0,6612	0,5918	0,5153	0,4551	0,1899	0,4844
15	94,58	0,6612	0,5980	0,5143	0,4524	0,1895	0,4819

Tabla A.8. Resultados sistema IR-n con diferentes tamaños de pasajes, subintervalo óptimo. Colección EFE con preguntas largas

		Precisión a los N documentos						
Sol.	**Cob.**	**5**	**10**	**20**	**30**	**200**	**AvgP**	**Tiempo**
LATimes								
1	95,09	0,4936	0,4149	0,3138	0,2589	0,0753	0,4744	3:27.53
2	94,39	0,4936	0,4064	0,3074	0,2518	0,0730	0,4649	3:20.30
4	93,57	0,5021	0,4000	0,2968	0,2518	0,0712	0,4513	3,19.85
8	93,11	0,4894	0,3894	0,2872	0,2454	0,0711	0,4477	3:19.52
EFE								
1	95,06	0,6694	0,5898	0,4990	0,4463	0,1940	0,5093	2:08.97
2	94,88	0,6776	0,5816	0,4949	0,4456	0,1930	0,5070	2:07.56
4	94,62	0,6857	0,5816	0,5010	0,4490	0,1910	0,5057	2:07.75
8	94,54	0,6816	0,5776	0,5010	0,4483	0,1901	0,5040	2:07.61

Tabla A.9. Utilización de diferentes grados de solapamiento. Preguntas cortas

		Precisión a los N documentos						
Sol.	**Cob.**	**5**	**10**	**20**	**30**	**200**	**AvgP**	**Tiempo**
LATimes								
1	95,44	0,5149	0,4191	0,3234	0,2652	0,0768	0,5025	14:29.76
2	94,98	0,5149	0,4128	0,3191	0,2674	0,0761	0,4973	14:00.54
4	94,98	0,4936	0,3872	0,3106	0,2553	0,0741	0,4548	13:48.22
8	94,51	0,4851	0,3936	0,3043	0,2525	0,0728	0,4598	13:42.62
EFE								
1	95,92	0,6857	0,6061	0,5041	0,4517	0,1961	0,5192	5:40.82
2	95,92	0,6857	0,6041	0,5010	0,4503	0,1953	0,5151	5:29.37
4	95,69	0,6939	0,6102	0,5000	0,4503	0,1934	0,5147	5:27.30
8	95,81	0,6939	0,6061	0,5000	0,4531	0,1933	0,5135	5:26.30

Tabla A.10. Utilización de diferentes grados de solapamiento. Preguntas largas

Volver al índice/Tornar a l'índex

		Precisión a los N documentos						
α	Cob.	5	10	20	30	200	AvgP	Inc
Preguntas cortas								
1	95,21	0,5106	0,4170	0,3096	0,2624	0,0752	0,4896	0
1,1	95,09	0,5319	0,4170	0,3117	0,2603	0,0751	**0,4927**	+0,6%
1,2	95,21	0,5277	0,4106	0,3149	0,2589	0,0750	0,4843	-1,1%
1,3	95,21	**0,5319**	0,4128	0,3149	0,2560	0,0749	0,4826	-1,4%
1,4	95,21	0,5277	0,4149	0,3128	0,2539	0,0749	0,4825	-1,5%
Preguntas largas								
1	95,79	0,5404	0,4106	0,3160	0,2674	0,0769	0,5188	0
1,1	95,91	0,5362	0,4128	0,3223	0,2674	0,0773	0,5204	+0,3%
1,2	95,68	0,5362	0,4128	0,3245	0,2645	0,0774	**0,5232**	+0,8%
1,3	95,68	0,5447	0,4085	0,3255	0,2652	0,0770	0,5105	-1,5%
1,4	95,56	0,5447	0,4043	0,3202	0,2610	0,0768	0,5114	-1,4%

Tabla A.11. Resultados aplicación medidas de proximidad. Colección LATimes

		Precisión a los N documentos						
α	Cob.	5	10	20	30	200	AvgP	Inc
Preguntas cortas								
1	95,06	0,6694	0,5898	0,4990	0,4463	0,1940	0,5093	0
1,1	95,25	0,6653	0,6000	0,5010	0,4503	0,1952	0,5114	+0,4%
1,2	95,25	0,6816	0,6000	0,5031	0,4497	0,1952	0,5109	+0,3%
1,3	95,10	0,6898	0,5980	0,5051	0,4503	0,1950	**0,5130**	+0,7%
1,4	94,95	0,6816	0,6020	0,5020	0,4517	0,1950	0,5117	+0,5%
Preguntas largas								
1	96,18	0,6776	0,5939	0,5061	0,4558	0,1968	0,5205	0
1,1	96,36	0,6694	0,5918	0,5061	0,4558	0,1981	**0,5227**	+0,4%
1,2	96,14	0,6653	0,5959	0,5041	0,4565	0,1992	0,5218	+0,2%
1,3	95,84	0,6653	0,5980	0,4990	0,4558	0,1990	0,5194	-0,2%
1,4	95,73	0,6612	0,5939	0,4959	0,4537	0,1983	0,5163	-0,8%

Tabla A.12. Resultados aplicación medidas de proximidad. Colección EFE