Método de Fusión Dinámica aplicado a la Recuperación de Información

por

Antonio Juárez González

M.C., INAOE

Tesis sometida como requisito parcial para obtener el grado de

Doctor en Ciencias en la Especialidad de Ciencias Computacionales

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica

Noviembre 2010

Tonantzintla, Puebla

Supervisada por:

Dr. Luis Villaseñor Pineda, INAOE

Dr. Manuel Montes y Gómez, INAOE

Resumen

El desarrollo de múltiples sistemas de Recuperación de Información (RI) ha permitido el buen aprovechamiento de la inmensa cantidad de información disponible en nuestros días. Sin embargo, no existe un sistema de RI que satisfaga todas las necesidades de información (*peticiones*) de los usuarios, sino que algunos sistemas obtienen buenos resultados para un tipo de peticiones mientras que para otro tipo de peticiones no son capaces de encontrar elementos relevantes. Esta situación ofrece un escenario donde se cuenta con difererentes resultados de recuperación, obtenidos con diferentes sistemas de RI, para una misma petición. El problema en esta situación es que se necesita entregar una sola lista de resultados al usuario. La estrategia más utilizada en este tipo de escenarios es la Fusión de Datos, la cual tiene como objetivo generar una nueva lista de resultados a partir de un conjunto de listas de resultados iniciales. Esta nueva lista debe ofrecer al usuario mejores resultados de recuperación (debe contener más elementos relevantes y/o colocarlos en las primeras posiciones de la lista) que la mayoría de las listas de resultados iniciales, preferentemente que todas ellas. Normalmente, la Fusión de Datos se utiliza sistemáticamente, es decir, se fusionan todas las listas disponibles para una petición. Sin embargo, estudios acerca de la efectividad particular de la fusión en cada petición, han revelado que en la mayoría de las veces no es apropiado fusionar todas las listas disponibles, y tampoco las listas con la mejor eficacia global. Por otro lado, la Fusión de Datos es sensible a la inclusión de listas con resultados de recuperación pobres, las cuales provocan un decremento en la efectividad de los métodos de fusión, dando como resultado una lista en ocasiones peor que cualquiera de las listas iniciales. En esta investigación se desarrolla un método que contribuye a mejorar la eficacia de los métodos de Fusión de Datos mediante la selección previa de las listas que serán incluídas en la fusión. Nuestro método,

llamado *Fusión Dinámica de Resultados de RI* (FDRI) se basa en las características de redundancia y posicionamiento de los elementos en las listas de elementos recuperados, y no requiere información acerca de los sistemas con los que se realizó el proceso de recuperación. Más aún, FDRI implementa un enfoque no supervisado, lo cual lo hace independiente de los juicios de relevancia y evita el reentrenamiento al ser aplicado a diferentes conjuntos de datos. Además, FDRI evita la generación de todas las posibles fusiones de las listas iniciales y es independiente del método de fusión utilizado. Los resultados experimentales obtenidos muestran que la aplicación del método FDRI antes del proceso de fusión mejora sustancialmente los resultados de la fusión sistemática considerando tres métodos de fusión: MaxRSV, CombMNZ y Fuzzy Borda. Los resultados obtenidos fueron validados estadísticamente mediante el método *paired Student's t-test*, mostrando que la mejora lograda por el método FDRI sobre los resultados de la fusión sistemática es estadísticamente significativa con una confianza del 95 %.

Abstract

The development of Information Retrieval Systems (IRS) has allowed the exploitation of the huge amount of information available nowadays. However, no IR system can satisfy all the information needs (*queries*) of the users. Instead, some IRSs are effective for some kind of queries while for some other kind of queries the same IRSs perform poorly. This situation leads to a scenario where, for the same query, there are available multiple retrieval result lists with different effectiveness for satisfying the information need. The problem in this situation is to decide which result list must be given to the user. The most used strategy to solve this problem is known as Data Fusion. The main goal of Data Fusion is the generation of a new list of retrieval results, constructed from the initial result lists. This new list must be more effective to satisfy the information need of the user (i.e., it must contain more relevant elements and/or rank the relevant elements better in the list) than the majority of the initial retrieved lists, preferably than all of them. Data Fusion is usually applied systematically, this is, all the available result lists for a query are fused. Nevertheless, studies about the effectiveness of Data Fusion by query show that, in most of the cases, to fuse all the available result lists, or to fuse the best global ones is not the best option. On the other hand, Data Fusion is highly sensitive to the inclusion of retrieval results with low effectiveness, which causes a decrease in the effectiveness of the fusion methods. This situation leads to a fusion result in some cases worse than all the initial retrieval results. In this research, a method that helps to increase the effectiveness of the Data Fusion methods by means of a previous selection of the result lists to be included in the fusion process, is developed. Our method, called *Dynamic Fusion of Information Retrieval Results* (DFIR), is based on the features of redundancy and ranking of the elements in the retrieval result lists, and does not require any information about the inner process of the IRSs used to retrieve the lists. Even more, DFIR is implemented in an unsupervised approach avoiding the dependence of the relevance judgments and

a possible re-training process if the data sets or the IRSs change. Also, DFIR avoids the generation of all the possible fusions of the initial result lists, and is not linked to a single Data Fusion method. Experimental results show that the use of the method DFIR before the fusion process substantially improves the effectiveness of the Data Fusion applied systematically, considering three different fusion methods: MaxRSV, CombMNZ and Fuzzy Borda. The *paired Student's t-test* method was used to validate all the obtained results, showing that the increase in the effectiveness achieved by the DFIR method, over the systematic Data Fusion, is statistically significant considering an error threshold of 0.5 %.

Índice General

Lista de Figuras

Lista de Tablas

Capítulo 1

Introducción

En la rutina diaria de la sociedad actual, podemos encontrar una tarea en común: la consulta de información. Si bien anteriormente bibliotecas, hemerotecas, museos, exposiciones, etc., eran la manera más adecuada para conseguir información, en esta época los avances tecnológicos han permitido recopilar grandes volúmenes de información, de prácticamente cualquier tema, en repositorios que pueden ser consultados mediante una computadora y una conexión a Internet. La llamada *Web* contiene actualmente la mayor cantidad de información disponible, que puede ser muy especializada, como los avances más recientes en medicina, o muy informal, como las vacaciones de una persona ansiosa por compartir con el mundo sus aventuras en la playa. De esta manera es como actualmente tenemos a la mano una herramienta para estar informados de las cosas que suceden en el mundo desde la comodidad de nuestra casa.

Desde su creación en 1969, donde sólo eran 4 máquinas interconectadas [25], la Internet, conocida actualmente como *WWW (World Wide Web* o simplemente *Web*), se ha visto beneficiada por diferentes avances tecnológicos como microprocesadores más rápidos y poderosos, buses de transferencia de datos más veloces y dispositivos de almacenamiento de datos de gran capacidad. Es este último avance tecnológico el que ha permitido mantener un registro de la información generada desde que las computadoras fueron creadas. Hoy en día la capacidad estándar de los discos duros comerciales va de los 160 gigabytes (1 gigabyte = 1024 megabytes) hasta 1 terabyte (1 terabyte = 1024 gigabytes), con un costo relativamente bajo. En un terabyte puede almacenarse una décima parte de la colección impresa de la Biblioteca del Congreso de Estados Unidos (la cual consta de 130 millones de elementos en cerca de 530 millas de estantes, que incluyen 29 millones de libros, 2.7 millones de grabaciones, 12 millones

de fotografías, 4.8 millones de mapas y 58 millones de manuscritos)[1]. Se estima que este año, 2010, habrá 988 exabytes (1 exabyte = 1024 petabytes; 1 petabyte = 1024 terabytes), poco menos de un zetabyte (1024 exabytes), en todas las computadoras de almacenamiento en todo el mundo [24]. Más aún, IBM calcula que después de 2010 el volumen de datos accesibles en línea, ya sea en Internet o en redes corporativas será de casi un yotabyte (1024 zetabytes) [2].

Todo el universo de información con la que se cuenta hoy en día ofrece una gran ventaja para quienes requieran conocimiento específico, ya sean estudiantes de todos los niveles académicos, investigadores, científicos, o simplemente personas que planean sus vacaciones. En nuestros días es muy común ingresar a un sitio de internet y utilizar el campo que tiene la etiqueta *Búsqueda* para ingresar ciertas palabras de la información que necesitamos y obtener una lista de aquellos elementos relacionados con el texto introducido. Imaginemos un escenario donde tuviéramos toda la información del mundo disponible pero no contáramos con esa herramienta de búsqueda mencionada anteriormente. El resultado sería catastrófico, sería como tener un almacén con billones y billones de libros, revistas, imágenes, videos y audios, y, al requerir alguna información específica, tener que consultar sección por sección, tipo por tipo y elemento por elemento, todo el contenido del almacén hasta encontrar aquello que nos interesa. En estas circunstancias, la información que pudiéramos extraer sería incompleta y tal vez errónea, por lo que la gran mayoría de la información disponible resultaría inútil.

Las Ciencias Computacionales han atacado este problema desde hace ya más de 40 años, mediante un área llamada *Recuperación de Información* (RI). Salton [29] en 1968 define esta área de investigación de la siguiente manera:

> *Recuperación de Información es un campo concerniente a la estructuración, análisis, organización, almacenamiento, búsqueda y recuperación de información.*

Por otro lado, la tarea de Recuperación de Información consiste en, dada una consulta formulada en lenguaje natural por algún usuario, obtener elementos relevantes que satisfagan las necesidades de información de dicho usuario [15].

De la problemática descrita anteriormente y de las definiciones del área y tarea de Recuperación de Información, han surgido multiples sistemas de RI que actualmente

[1] http://www.sentientdevelopments.com/2007/03/whole-lotta-bits-n-bytes.html
[2] http://e-rgonomic.blogspot.com/2008/04/la-sociedad-overload.html

sirven como herramientas para consultar y recuperar documentos, imágenes, videos o audio de diferentes colecciones disponibles de manera gratuita en la Web, o colecciones restringidas pertenecientes a empresas o instituciones.

Un sistema de RI recibe una petición formulada en lenguaje natural y ofrece como resultado una lista de elementos relacionados con la petición, ordenados de acuerdo a una medida de similitud. Se espera que la lista resultante presente en las primeras posiciones los elementos más relevantes a la petición, y los menos relevantes en las últimas posiciones.

Los ejemplos más representativos de un sistema de RI, además de ser los más utilizados actualmente, son los llamados *Buscadores de Internet*. Entre los más conocidos podemos mencionar a Google (*www.google.com*), Yahoo (*www.yahoo.com*), MSN (*www.msn.com*), Lycos (*www.lycos.com*), Altavista (*www.altavista.com*) y Excite (*www.excite.com*). Sin embargo, la utilidad de los sistemas de RI no se limita a actividad en la Web, ya que existen colecciones de elementos no disponibles en internet que presentan la misma problemática: mucha información imposible de revisar manualmente. Para estas situaciones existen aplicaciones que permiten realizar todo el proceso de un sistema de RI: representación, almacenamiento, organización y acceso a la información [48]. Lemur (*www.lemurproject.org*), Lucene (*lucene.apache.org*) y Google Desktop (*desktop.google.com*), son ejemplos de estas aplicaciones.

Dentro del proceso de desarrollo de los sistemas de RI se requiere evaluar la eficacia de los mismos, para lo cual existen conferencias y foros de evaluación que se han llevado a cabo con este fin, siendo las más representativas la *TREC (Text REtrieval Conference)* para el idioma Inglés y el *CLEF (Cross Language Evaluation Forum)* para idiomas europeos. En estos eventos se utilizan colecciones estándar que permiten comparar de una manera justa y controlada, la eficacia de los diferentes sistemas de RI participantes. Además, estos eventos permiten que grupos de trabajo de todo el mundo comparten sus investigaciones y avances en el área de RI.

La gran cantidad y variedad de información disponible, y la investigación en el área de RI, ha dado como resultado una gran cantidad de métodos de RI cuyas diferencias radican principalmente en la forma en que se representan internamente los elementos de la colección y/o en la medida de similitud utilizada para determinar los elementos relevantes a una petición. Otra forma de explicar la continua investigación en RI es el hecho de que no existe un sistema que satisfaga todas las necesidades de un usuario; en general, algunos sistemas son *buenos* encontrando información relevante

para ciertas peticiones pero son *malos* para otras [3]. Una manera de solucionar esta situación es tratar de desarrollar un buscador que tome en cuenta todos los posibles tipos de peticiones y colecciones de búsqueda que puedan presentarse, pero debido a la gran variedad de temas, estilos, estructuras, niveles y formatos en que puede generarse la información digital, y las inmensas posibilidades que tienen los usuarios de escribir su petición en lenguaje natural, esto resulta prácticamente imposible. Una mejor opción es aprovechar los resultados de diferentes sistemas de RI (SRIs) para una misma petición, y combinarlos para tratar de obtener una mejor lista de resultados. De esta manera se intenta aprovechar los buenos resultados de ciertos sistemas para algunas peticiones y se intenta eliminar las flaquezas que presentan los sistemas para otras peticiones. Este último enfoque para mejorar los resultados de la recuperación recibe el nombre de Fusión de Datos (FD).

La Fusión de Datos es un proceso para combinar información extraída por diferentes sistemas en un solo resultado final [16]. Lo que se espera del resultado de la Fusión de Datos, en el área de Recuperación de Información, es que sea más eficaz al satisfacer a la necesidad de información que la mayoría de los resultados extraídos por los diferentes SRIs, preferentemente que todos. En una situación cotidiana donde no se cuenta con información acerca de la eficacia de los resultados de los SRIs y se requiere ofrecer un sólo resultado al usuario, la Fusión de Datos debe garantizar un resultado mejor que la mayoría de los SRIs.

Este enfoque a sido ampliamente utilizado con el fin de mejorar los resultados de recuperación para una petición utilizando un conjunto de resultados previamente extraídos con diferentes sistemas de RI [6, 10, 12, 14, 21, 23, 26, 28, 32, 34, 35, 36, 38, 39, 40, 41, 43, 46]. Los trabajos citados muestran que es posible mejorar, globalmente, los resultados de recuperación al aplicar métodos de Fusión de Datos. Sin embargo, algunos de estos trabajos también muestran que, de manera individual, para muchas peticiones no se logra el objetivo de mejorar la eficacia de la mayoría de las listas iniciales. Este hecho indica que, si bien la FD mejora en gran magnitud los resultados de recuperación de algunas peticiones, también tiene el problema de decrementar mucho su eficacia en otras. Existen diferentes métodos para combinar resultados, los cuales pueden ser agrupados en: *lineales*, basados en los *scores* de los elementos de las listas; *posicionales*, basados en las posiciones de los elementos en las

[3] Análisis de los resultados de diferentes sistemas de RI pueden consultarse en la sección de publicaciones y working notes TREC (*trec.nis.gov*) y CLEF (*www.clef-campaign.org*).

listas; *probabilísticos*, basados en la probabilidad de encontrar elementos relevantes en segmentos de la lista de resultados; y los basados en la *teoría de elección social*, donde cada lista de resultados se toma como un *votante*, cada elemento en las listas es un *candidato* y el *score* de cada elemento es la confianza del votante hacia el candidato.

Aunque la Fusión de Datos ha demostrado ser de gran utilidad cuando se cuenta con diferentes resultados para una misma petición, la elección de las listas que deben ser combinadas y el modo en que deben combinarse sigue siendo un problema abierto. Esta investigación aborda esta problemática, haciendo uso de características propias de las listas de resultados, evitando dependencias del método de fusión a utilizar, de los *juicios de relevancia*[4], de entrenamiento previo, del funcionamiento interno de los sistemas de recuperación y de la generación de todas las posibles fusiones.

1.1. Descripción del problema

La aplicación de la Recuperación de Información necesita de dos elementos: una gran cantidad de información, a la cual llamaremos *colección*, y una necesidad de información, a la cual llamaremos *petición*. Los elementos en la colección deben reducirse a una representación manejable por los SRIs, la cual recibe el nombre de *índice*. Este índice es utilizado para identificar los elementos relevantes a la petición, de acuerdo a distintas medidas de similitud [5]. Con lo anterior, el proceso de Recuperación de Información toma una petición en lenguaje natural, la cual es reducida a la representación interna de los elementos de la colección; un SRI realiza un proceso de comparación entre la petición y los elementos de la colección obteniendo una lista ordenada de elementos relacionados con la petición.

La eficacia de los sistemas de RI para satisfacer las necesidades de información (peticiones), normalmente se obtiene aplicando medidas de evaluación a nivel global [8, 19, 22, 52, 54]. Lo anterior significa que, dado un conjunto de peticiones de prueba y una colección de búsqueda, la eficacia global del sistema se determina por el promedio de la eficacia obtenida en cada una de las peticiones.

Con el creciente aumento de información y la variedad de la misma (documentos, imágenes, video y audio), las investigaciones en RI han propiciado modificaciones al

[4]Los juicios de relevancia son una lista de los elementos relevantes existentes para las peticiones en la colección de búsqueda. Se construyen de manera manual y sólo están disponibles para colecciones de evaluación (ver sección 2.1.4).

[5]Detalles del proceso de RI se presentan en el capítulo 2

esquema básico. Las modificaciones más utilizadas se basan en múltiples formulaciones de la petición procesadas por un único sistema de RI, y la utilización de múltiples sistemas de RI para procesar una misma formulación de la petición. Otra posibilidad, aunque no muy investigada actualmente, es una combinación de las dos modificaciones anteriores, es decir, tener múltiples formulaciones de la petición y procesarlas con múltiples sistemas de RI. Algo común en los tres procesos modificados de RI es que la salida de estos procesos son diferentes listas de resultados.

Dado que se debe entregar una sola lista de resultados al usuario para satisfacer su necesidad de información, estos esquemas deben decidir qué lista de resultados arrojadas por su proceso de IR debe entregarse. Usualmente, la decisión se toma con base en la observación de la eficacia de las distintas configuraciones, ya sea formulación, sistema o formulación-sistema. La lista de resultados seleccionada es aquella que obtiene los mejores resultados de recuperación en un conjunto de peticiones de prueba [2, 3, 7, 17, 18, 51, 56].

Como se comentó anteriormente, una opción para aprovechar múltiples resultados de recuperación para una misma petición, obtenidos mediante diferentes reformulaciones y/o sistemas de RI, es la Fusión de Datos. Esta estrategia genera una nueva lista de resultados al combinar, mediante diferentes procedimientos, los resultados de recuperación disponibles. El proceso de Fusión de Datos se muestra en la figura 1.1.

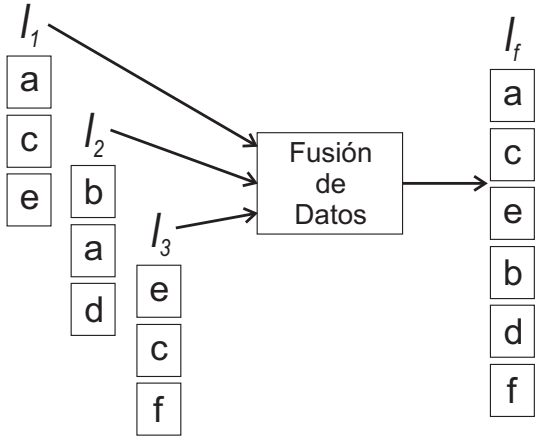

Figura 1.1: Proceso de Fusión de Datos

En la figura 1.1 podemos observar dos características necesarias para la Fusión de Datos: un conjunto de listas (l_1, l_2 y l_3) y elementos repetidos en las listas (a, c y e). En general, los métodos de Fusión de Datos se basan en la redundancia y posicionmiento de los elementos en las listas para determinar a los elementos con más

probabilidad de ser relevantes, y formar con ellos una nueva lista de resultados de recuperación [6].

La evaluación de la eficacia de los métodos de fusión es calculada de manera similar a como se hace con los sistemas de RI: para cada petición se determina la eficacia que la lista resultante de la FD tiene para satisfacer la necesidad de información, y se realiza un promedio para conocer la eficacia global. Esta situación suele beneficiar a sistemas de RI o métodos de fusión que logran mejorar, en gran medida, los resultados de recuperación de algunas peticiones pero disminuyen su eficacia en muchas otras. La tabla 1.1 muestra un ejemplo de lo anterior.

Petición	$\mathbf{S}_b$	$\mathbf{S}_w$	**Promedio**	$\mathbf{FD}_1$	$\mathbf{FD}_2$
$\mathbf{p}_2$	0.076	0.054	0.048	**0.105**	0.059
$\mathbf{p}_3$	**0.226**	0.030	0.104	0.085	0.089
$\mathbf{p}_5$	0.209	0.072	0.149	**0.230**	0.229
$\mathbf{p}_6$	0.454	0.065	0.370	0.443	**0.543**
$\mathbf{p}_{10}$	0.176	0.004	0.100	**0.191**	0.156
$\mathbf{p}_{15}$	0.528	0.153	0.539	0.555	**0.704**
MAP	0.278	0.063	- -	0.268	**0.297**

Tabla 1.1: Análisis del AP y MAP de sistemas de RI y la FD para diferentes peticiones

Para seis peticiones se obtuvieron cinco resultados de recuperación mediante diferentes estrategias, de las cuales en la tabla 1.1 se muestra la eficacia de la mejor ($\mathbf{S}_b$) y de la peor ($\mathbf{S}_w$) de ellas, considerando la medida de evaluación AP[7]. Esta tabla también muestra el promedio de las eficacias obtenidas por las cinco estrategias de recuperación para cada petición (**Promedio**), y el AP de dos métodos de fusión distintos ($\mathbf{FD}_1$ y $\mathbf{FD}_2$). En negritas se presentan los valores más altos.

Podemos observar que el método $\mathbf{FD}_2$ obtiene el mejor MAP[8], sin embargo, su eficacia individual sólo es superior a la del método $\mathbf{FD}_1$ en dos de las seis peticiones ($\mathbf{p}_6$ y $\mathbf{p}_{15}$). Por otro lado, $\mathbf{S}_b$ también tiene un mejor MAP que $\mathbf{FD}_1$, pero sólo es mejor individualmente en dos peticiones. En este ejemplo, el mejor resultado debería ser $\mathbf{FD}_1$ ya que ofrecería más elementos relevantes y/o mejor posicionados que los otros dos resultados, sin embargo, considerando el MAP este resultado es el peor.

Otro aspecto importante que debemos notar de los resultados individuales obtenidos por los métodos de fusión, es que en casi todos las peticiones se obtiene un

[6] Detalles de diferentes métodos de fusión se muestran en el capítulo 2

[7] *Average Precision*. Detalles de esta medida se presentan en el capítulo 2.

[8] *Mean Average Precision*. Detalles de esta medida se presentan en el capítulo 2.

AP mayor al promedio de los AP de las listas iniciales para cada petición (columna **Promedio**). El único caso donde esto no sucede es para la petición $\mathbf{p}_3$ con ambos métodos de fusión, sin embargo, los AP de los métodos de fusión son mayores al de la peor lista inicial ($\mathbf{S}_w$) para esta misma petición.

Del ejemplo anterior podemos hacer las siguientes observaciones:

1. La fusión no siempre logra mejores resultados que los obtenidos por la mejor de las listas iniciales, pero generalmente es mejor que la peor de ellas.

2. Una mejora en el MAP no es garantía de que todos los resultados mejoraron individualmente, ya que puede suceder que sólo se logre mejorar en gran medida la eficacia de algunas peticiones.

3. El hecho de que la fusión mejore al promedio de los AP de las listas iniciales, y de que su eficacia sea equivalente a la de la mejor lista inicial ($\mathbf{S}_b$) en la mayoría de los casos, muestra indicios de que la fusión mejora a la mayoría de las listas iniciales.

Considerando los resultados de la tabla 1.1 y las observaciones anteriores, podemos notar que el método $\mathbf{FD}_1$ no logra superar a la mejor lista fusionada en dos peticiones ($\mathbf{p}_3$ y $\mathbf{p}_6$), mientras que el método $\mathbf{FD}_2$ no supera a la mejor lista en tres peticiones ($\mathbf{p}_2$, $\mathbf{p}_3$ y $\mathbf{p}_{10}$). Esta situación es muy común cuando se aplican métodos de fusión.

Veamos un ejemplo más a fondo. Consideremos el AP de los resultados de cinco diferentes sistemas de RI para dos peticiones (tabla 1.2). Podemos observar que para la petición $\mathbf{p}_a$ el sistema $\mathbf{S}_2$ es el mejor, mientras que para la petición $\mathbf{p}_b$ el mejor sistema es $\mathbf{S}_1$. Esta tabla nos muestra la variabilidad de resultados de los sistemas de RI para una misma petición, así como su variabilidad al tratar diferentes peticiones.

Ahora consideremos los APs de las posibles fusiones de dos resultados, y el de la fusión de los resultados de los cinco SRIs (tabla 1.3).

Los datos de la tabla 1.3 fueron obtenidos fusionando los resultados de recuperación de los cinco SRIs mediante el método CombMNZ [9]. Esta tabla muestra que para la petición $\mathbf{p}_a$ ninguna fusión logra superar al mejor resultado individual (tabla 1.2). Por otro lado, para la petición $\mathbf{p}_b$, ocho fusiones de las once generadas logran superar al mejor resultado individual. Sin embargo, la mejor fusión ($\mathbf{S}_2$-$\mathbf{S}_4$) no incluye al mejor sistema ($\mathbf{S}_1$). Este ejemplo muestra evidencia de que, si bien la

[9] Detalles de este método de fusión se muestran en la sección 2.2.2

p_a	**military intervention of Russia in Chechenya**
Sistema	***AP***
S_1	0.629
S_2	**0.661**
S_3	0.291
S_4	0.485
S_5	0.470
p_b	**invasion of Haiti by U.N./US soldiers**
Sistema	***AP***
S_1	**0.256**
S_2	0.221
S_3	0.162
S_4	0.245
S_5	0.251

Tabla 1.2: Eficacia de diferentes sistemas de RI para dos peticiones

p_a		p_b	
Fusión	***AP***	**Fusión**	***AP***
S_1-S_2	0.655	S_1-S_2	0.245
S_1-S_3	0.579	S_1-S_3	0.255
S_1-S_4	0.599	S_1-S_4	**0.284**
S_1-S_5	0.614	S_1-S_5	**0.290**
S_2-S_3	0.591	S_2-S_3	0.248
S_2-S_4	0.621	S_2-S_4	***0.294***
S_2-S_5	0.639	S_2-S_5	**0.290**
S_3-S_4	0.513	S_3-S_4	**0.256**
S_3-S_5	0.562	S_3-S_5	**0.269**
S_4-S_5	0.510	S_4-S_5	**0.257**
S_1-S_2-S_3-S_4-S_5	0.615	S_1-S_2-S_3-S_4-S_5	**0.285**

Tabla 1.3: Eficacia de diferentes fusiones para dos peticiones.

Fusión de Datos es una buena opción para mejorar los resultados de la Recuperación de Información, la baja calidad de algunas de las listas fusionadas influyen en gran medida en la calidad del resultado final.

De lo anterior podemos hacer las siguientes observaciones:

1. Para algunas peticiones la Fusión de Datos no logra superar la eficacia de la mejor lista individual.

2. Cuando se generan fusiones de subconjuntos de las listas disponibles, suele

suceder que alguna fusión logre superar a la mejor lista individual.

3. En algunas ocasiones, la fusión más eficaz para satisfacer la necesidad de información no contiene a la mejor lista individual.

4. La fusión utilizada sistemáticamente (fusionar todas las listas disponibles) no siempre es la mejor opción.

Con todo lo anterior, la problemática en el ámbito de Fusión de Datos que esta investigación aborda se resume en las siguientes preguntas:

1. ¿Bajo qué condiciones la Fusión de Datos sistemática no logra mejorar los resultados de RI?

2. ¿En qué medida una fusión no sistemática, aplicada particularmente a los resultados de cada petición, puede ayudar a mejorar los resultados de RI?

3. ¿Cómo puede medirse la utilidad de una lista para ser incluida en la fusión?

Las interrogantes anteriores hoy en día son un problema abierto, y, en nuestro conocimiento, existen sólo unos cuantos trabajos que la abordan a manera de análisis. Esta investigación va un paso más allá al proponer un método que determina las listas que deben ser incluidas en el proceso de Fusión de Datos, y lo aplica directamente al proceso de RI. Más aún, el método propuesto es aplicado a cada petición de manera particular, logrando identificar diferentes subconjuntos de listas a fusionar en cada caso. Esta característica es la que aporta el dinamismo a nuestro método, al cual llamamos Fusión Dinámica de Resultados de RI (FDRI).

La sección siguiente describe la solución propuesta para abordar la problemática, así como un diagrama que muestra de manera general el método de Fusión Dinámica de Resultados.

1.2. Solución propuesta

La problemática presentada involucra una predicción de la eficacia que tendrá la Fusión de Datos al ser aplicada a un conjunto de listas predefinidas. Este problema se ha tratado mediante un enfoque de Aprendizaje Automático (AA) [9, 16, 37, 44, 55]. Estos estudios han demostrado la factibilidad de conocer la eficacia que tendrá una

fusión, dadas ciertas características de las listas que serán fusionadas y/o de los SRIs utilizados para recuperarlas. Sin embargo, debido a que este enfoque es por naturaleza supervisado, son necesarios los *juicios de relevancia*[10] para poder realizar el entrenamiento. Otra limitante de algunos de estos enfoques es que necesitan realizar las posibles fusiones de distintos subconjuntos de las listas disponibles para realizar su entrenamiento. Si bien estos trabajos muestran la factibilidad de abordar el problema, no ofrecen una forma de aplicar sus resultados al proceso de Recuperación de Información.

En este trabajo de tesis proponemos un método para realizar un análisis previo al proceso de Fusión de Datos, para seleccionar las listas que deben ser fusionadas y ofrecer una lista de resultados al usuario, lo cual permite incluirlo de manera natural en el proceso de RI. La

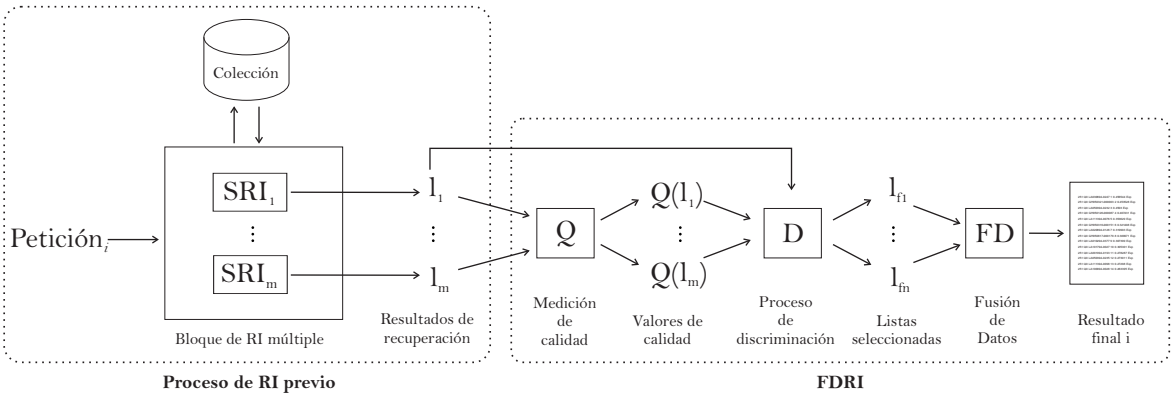

Figura 1.2: Esquema general del método FDRI para la petición i

El método propuesto recibe como entrada un conjunto de resultados de recuperación ($\mathbf{l}_1$, ..., $\mathbf{l}_m$). La calidad de cada resultado es obtenida aplicando una medida de calidad ($\mathbf{Q}$). Una vez calificadas, las mejores n listas, de acuerdo a un proceso de discriminación, son seleccionadas para el proceso de Fusión de Datos. La lista fusionada es ofrecida como resultado final al usuario. Lo anterior se realiza para cada petición, por lo que las listas que se fusionan no siempre son las mismas. El método propuesto, al cual llamamos Fusión Dinámica de Resultados, tiene como componentes principales la medida de calidad $\mathbf{Q}$ y el proceso de discriminación $\mathbf{D}$, ya que con base

[10]Los juicios de relevancia son una lista de los elementos relevantes existentes para las peticiones en la colección de búsqueda. Se construyen de manera manual y sólo están disponibles para colecciones de evaluación (ver sección 2.1.4).

en éstos se seleccionan las listas que serán fusionadas cada vez. En el capítulo 5 se muestran los detalles de las diferentes medidas de calidad propuestas en esta investigación, así como los criterios de discriminación considerados.

1.3. Objetivos

En nuestro conocimiento, a la fecha no hay trabajos que realicen un análisis previo de las listas que serán fusionadas, ni que incluyan este proceso en la Recuperación de Información. Esta investigación constituye un esfuerzo para incluir la Fusión Dinámica de Resultados en el proceso de RI, persiguiendo los siguientes objetivos:

Objetivo general:

- Desarrollar un método automático de preselección de listas a fusionar (método FDRI) que mejore los resultados de la Fusión de Datos sistemática.

Objetivos específicos:

- Determinar características de las listas de resultados que permitan conocer su utilidad para el proceso de fusión, evitando la dependencia con los juicios de relevancia, los enfoques supervisados, el funcionamiento interno de los SRIs y la generación a priori de todas las posibles fusiones de las listas de resultados iniciales.
- Investigar la capacidad de características de las listas que no utilizan información del contenido de los elementos recuperados para identificar a las listas más aptas para la fusión.

1.4. Organización de la tesis

El resto del documento se organiza de la siguiente manera: el capítulo 2 describe los conceptos básicos de los dos temas principales de esta tesis, Recuperación de Información y Fusión de Datos; el capítulo 3 muestra los trabajos que han abordado el problema y muestra sus diferencias con el método propuesto; en el capítulo 5 se presentan los detalles de nuestro método de Fusion Dinámica de Resultados, así como diferentes estudios del mismo en los conjuntos de datos utilizados; por último, el

capítulo 7 presenta nuestras conclusiones y las posibles líneas de trabajo futuro que se desprenden de nuestros resultados.

Capítulo 2

Preliminares

La investigación presentada en este documento se relaciona con dos áreas principalmente: Recuperación de Información y Fusión de Datos. A continuación se detallan los conceptos básicos de ambas, para la correcta interpretación y entendimiento del método desarrollado en esta investigación, así como de los experimentos realizados y los resultados obtenidos.

2.1. Recuperación de información

Con el aumento diario de la información disponible en la Web, el contar con mecanismos de búsqueda y consulta de la misma se vuelve fundamental para poder satisfacer nuestras necesidades de información. El *Tratamiento Automático de Textos* (TAT) provee diferentes métodos para explorar documentos individuales o una colección de documentos en busca de información específica. Estos métodos son la *Extracción de Información*, la *Generación de Resúmenes*, la *Búsqueda de Respuestas* y la *Recuperación de Información*.

Cada uno de los métodos tiene una utilidad específica para ayudarnos a revisar, de una manera más sencilla, una cantidad mayor de documentos en un menor tiempo. Por ejemplo, la Extracción de Información (EI) ofrece una plantilla de conceptos previamente definidos como importantes, acerca de una temática específica; instancias de estos conceptos son extraídos de documentos que contengan información relacionada con la temática en cuestión. Si los documentos contienen información acerca de películas, los conceptos importantes podrían ser el director, los actores, el género, duración de la película y el país de origen. Este método evita que el usuario lea la totalidad de los documentos en busca de dichos conceptos.

Por otro lado, la Generación de Resúmenes (GR) ofrece al usuario un extracto de un documento, el cual contiene la información más relevante de acuerdo a los criterios particulares del método.

Los métodos de Búsqueda de Respuestas (BR) permiten al usuario ingresar una pregunta formulada en lenguaje natural y obtener una respuesta concreta a su pregunta. A diferencia de los métodos anteriores, los cuales actúan sobre documentos independientes, la BR actúa sobre una colección de documentos.

Por último, la Recuperación de Información (RI) permite al usuario expresar su necesidad de información ya sea con palabras clave, un enunciado o una descripción de varias líneas de texto. El resultado de estos métodos es una lista de documentos que contienen información relacionada con la necesidad de información del usuario (a la cual también se le llama *petición*), ordenados de acuerdo a un valor de similitud entre documento y petición. Es el usuario quien determina la utilidad de los documentos recuperados, sin embargo, la lista proporcionada por un método de recuperación evita al usuario la revisión exhaustiva de todos los documentos de la colección.

Aunque cada uno de los métodos mencionados anteriormente puede ayudarnos a acelerar la búsqueda y/o revisión de documentos, es la Recuperación de Información el más utilizado en nuestros días debido a la popularidad de su aplicación más conocida: los llamados *buscadores de internet*. Basados en un *Sistema de Recuperación de Información*, los buscadores ofrecen a los usuarios una manera transparente para consultar la inmensa cantidad de información contenida en la *Web*. Además de los buscadores, actualmente existe una gran variedad de sistemas de RI aplicados no sólo a la Web, sino también a colecciones no públicas de diferentes tipos de información, siendo la información textual la más tratada hasta ahora. La figura 2.1 muestra el esquema básico del proceso de RI.

En el esquema básico de Recuperación de Información, un SRI toma una petición en lenguaje natural, mediante un proceso ($\mathbf{R}$) la reduce a la representación interna de los elementos de la colección ($\mathbf{p}_R$). Después, la petición es comparada con los elementos de la colección de acuerdo a una medida de similitud ($\mathbf{S}$), con lo cual se obtiene una lista de resultados de recuperación que es entregada al usuario.

De manera general, los sistemas de RI pueden dividirse en dos categorías: los basados en técnicas de Procesamiento del Lenguaje Natural (PLN), y los estadísticos. Los primeros intentan implementar algún grado de "entendimiento" del lenguaje contenido en los documentos y la petición (por ejemplo, etiquetado de partes de

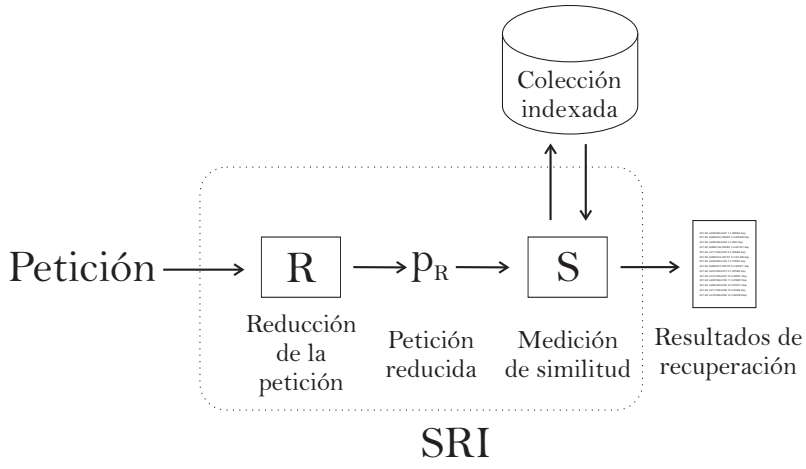

Figura 2.1: Esquema básico de Recuperación de Información

la oración, análisis sintáctico, detección de entidades nombradas,etc.). Los segundos tratan a los documentos y a la petición como conjuntos de palabras, siendo frecuencias y posicionamiento de estos la información utilizada para definir medidas de similitud que determinan a los documentos más relacionados con la petición del usuario. A la fecha, existe una mayor investigación sobre los métodos estadísticos, debido a su buen desempeño y relativa simplicidad. A continuación discutiremos sus características principales.

2.1.1. Representación interna de los documentos y la petición

Los métodos estadísticos reducen a los documentos y a la petición a un conjunto de palabras. Estas palabras son preprocesadas con el fin de eliminar la variación de la ocurrencia de diferentes formas gramaticales de la misma palabra, por ejemplo "utilizar", "utilizado" y "utilización". A este proceso se le llama *lematización*. Otro preprocesamiento común es la eliminación de las palabras que son comunes entre todos los documentos y por tanto no son útiles para diferenciar unos de otros. A este proceso se le llama *eliminación de palabras vacías*. Al final, las palabras conservadas después del preprocesamiento son utilizadas para generar la representación interna.

La mayoría de los sistemas de RI actuales implementan una representación de los documentos llamada *modelo de espacio vectorial*. En este modelo, los documentos son

codificados como vectores en un espacio k-dimensional. La elección de k puede basarse en el número de palabras, términos únicos, conceptos o tal vez clases asociadas con la colección de documentos [42]. Usualmente se utiliza el número de palabras diferentes, las cuales reciben el nombre de *vocabulario*. Cada componente del vector es utilizado para reflejar la importancia de la palabra correspondiente al representar el contenido de un documento.

Al representar a todos los documentos de la colección se genera una matriz, en la cual cada línea i representa un documento y cada columna j un palabra del vocabulario (figura 2.2).

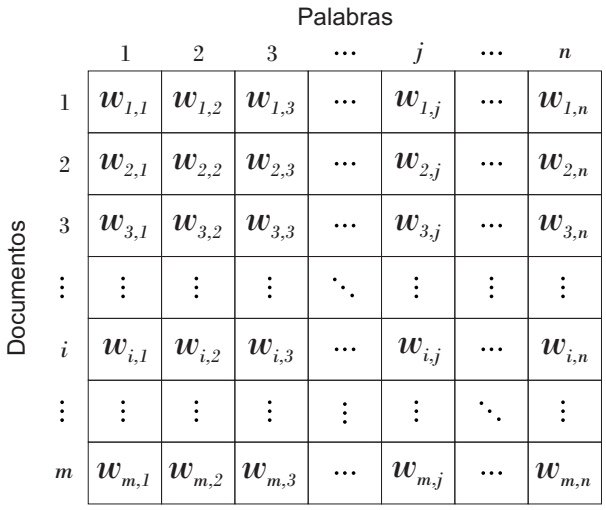

Figura 2.2: Modelo de espacio vectorial.

Cada entrada $w_{i,j}$ es un peso que representa el grado de importancia que tiene la palabra j para el documento i. La forma más sencilla de pesar a cada palabra es en una forma Booleana, la cual indica la existencia (1) o no existencia (0) de la palabra en el documento. Sin embargo hoy en día existen esquemas de pesado más representativos y útiles para discriminar la importancia entre documentos, dada una petición del usuario.

2.1.2. Esquema de pesado de palabras

El modelo de espacio vectorial define una forma robusta de representar a los documentos, sin embargo, los pesos Booleanos resultan ser insuficientes para discriminar la importancia entre los documentos de la colección, dada una petición del usuario. Por

lo anterior se han propuesto diferentes formas de pesar a las palabras para lograr una discriminación mayor, esto es, para poder determinar cuándo las palabras contenidas en un documento tienen más relación con la petición que los de otro. En general, los esquemas de pesado se basan en *frecuencias de palabras*, ya sea dentro del propio documento, dentro de la colección o una combinación de ambas.

De manera intuitiva, dada una petición, si un documento contiene las palabras de dicha petición en mayor frecuencia que otro, entonces dicho documento debería tener más relevancia para la petición. Lo anterior define un esquema de pesado llamado *term frequency* (*tf*). La frecuencia de la palabras es usualmente normalizada para evitar dar preferencia a documentos extensos. Formalmente, la importancia de la palabras j en documento i, de acuerdo al esquema *tf*, se define como:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,k}} \tag{2.1.1}$$

donde $n_{i,j}$ es el número de ocurrencias de la palabra j en el documento i, y el denominador es la suma del número de ocurrencias de todas las palabras en el documento i. *tf* es una medida local, ya que determina la importancia de una palabra dentro de un documento.

Por otro lado, para que una palabra pueda ser utilizada como un discriminante efectivo, ésta no debe estar presente en todos los documentos. En general, si una palabra existe sólo en un documento, esa palabra sería el mejor elemento para diferenciar a dicho documento del resto de la colección. Con esto en mente, se propuso una medida llamada *inverse document frequency* (*idf*), la cual es una medida global que indica qué tan distribuida está una palabra en los documentos de la colección, y por tanto qué tan probable es que aparezca dentro de un documento por azar. Formalmente *idf* se define como:

$$idf_j = log\Big(\frac{N}{|\{d : t_j \in d\}|}\Big) \tag{2.1.2}$$

donde N es el número de documentos en la colección y $|\{d : t_j \in d\}|$ es el número de documentos de la colección que contienen a la palabra t_j. De esta manera, cuanto menos aparezca una palabra en los documentos, *idf* es mayor. Si una palabra apareciera en todos los documentos, *idf* sería cero, es decir, dicha palabra no aporta ningún poder de discriminación entre documentos.

Hoy en día, una combinación de los esquemas presentados, llamada *tfidf*, es la más

ampliamente utilizada por sistemas de RI. Esta se define como:

$$tfidf_{i,j} = tf_{i,j} \times idf_j \tag{2.1.3}$$

Esta medida premia localmente a una palabra que aparece muy frecuentemente en un documento, pero la castiga si ésta aparece en muchos documentos de la colección.

Los anteriores esquemas de pesado, ofrecen una mejor forma de cuantificar la importancia de las palabras para discriminar a los documentos de la colección. Otros esquemas de pesado pueden consultarse en [31, 42].

2.1.3. Medida de similitud entre la petición y los documentos

Cuando un usuario introduce una petición a un sistema de RI, dicha petición es reducida a la representación de espacio vectorial, es decir, es representada como un vector de palabras. Este vector es utilizado para buscar a los documentos que tengan mayor relación con la petición, utilizando *medidas de similitud* entre vectores.

Una vez que se han representado los documentos y la petición como vectores, lo siguiente es determinar qué documentos son más relevantes a la petición, o dicho de otra forma, determinar cuáles *vectores-documento* son más parecidos al *vector-petición*. Para esto se utilizan *medidas de similitud*, siendo una de las más utilizadas la *similitud del coseno*. Esta medida se define como:

$$sim(q, d_i) = \frac{\sum_{j=1}^{t} w_{q,j} \times w_{i,j}}{\sqrt{\sum_{j=1}^{t} (w_{q,j})^2 \times \sum_{j=1}^{t} (w_{i,j})^2}} \tag{2.1.4}$$

donde q es el vector-petición, d_i es el vector que representa al documento i, $w_{q,j}$ es el peso de la palabra j del vector-petición y $w_{i,j}$ es el peso de la palabra j del vector-documento d_i. La mayor similitud que ofrece esta medida es 1, y se logra cuando el ángulo entre los vectores es 0, indicando que son paralelos. Lo anterior se interpreta como una completa identidad entre la petición y el documento. La menor similitud es 0, lo cual indica que los vectores son perpendiculares entre si(hay un ángulo de 90 grados entre ellos), y por tanto no tienen ninguna palabra en común. Diferentes medidas de similitud pueden consultarse en [31, 42].

La figura 2.3 muestra el proceso de RI tomando en cuenta cada uno de los componentes mencionados.

Como se dijo anteriormente, el resultado final de la Recuperación de Información es

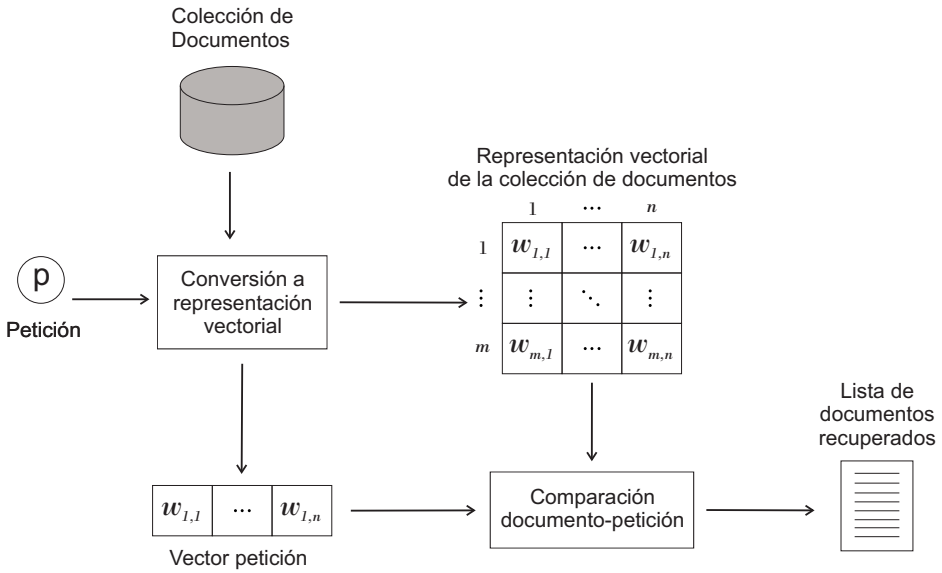

Figura 2.3: Proceso de Recuperación de Información extendido

una lista de documentos ordenados de acuerdo a su valor de similitud con la petición. Es el usuario quien determina la utilidad de los documentos recuperados. De manera intuitiva, si el usuario encuentra la información que busca dentro de los primeros documentos, entonces el sistema de RI obtuvo un buen resultado para la petición introducida. En los foros de evaluación para sistemas de RI, se utilizan diferentes estrategias para medir su eficacia al satisfacer necesidades de información. Algunas de las más utilizadas se presentan a continuación.

2.1.4. Evaluación

Para medir la eficacia de un sistema de RI al procesar una petición necesitamos dos elementos: una lista de documentos recuperados y una lista de los documentos dentro de la colección completa, que son relevantes a la petición. El primer elemento es el resultados de todo sistema de RI, sin embargo, el segundo elemento sólo puede obtenerse haciendo una revisión manual de los documentos de la colección para determinar aquellos que pueden satisfacer la necesidad de información descrita por la petición. Estos elementos reciben el nombre de *juicios de relevancia* (también pueden ser referidos como *documentos relevantes*, *juicios de evaluación*, *gold standard* o *ground truth*). Los juicios de relevancia son caros de obtener, en cuestión de tiempo y esfuerzo, por lo que para evaluar un sistema de RI en desarrollo suelen utilizarse con-

juntos de prueba estándar, dentro de foros de evaluación como el TREC o el CLEF, que se componen de una colección de documentos, un conjunto de peticiones y una conjunto de juicios de relevancia para cada petición. Una vez que se cuenta con los dos elementos mencionados, la evaluación observa dos aspectos: cuántos documentos relevantes se recuperaron y en que posición de la lista se encuentran.

La medida que toma en cuenta el número de documentos relevantes recuperados se llama *recuerdo* (*recall* en Inglés), denotada por la letra R, y se define como el número de documentos relevantes recuperados sobre el número de documentos relevantes existentes en la colección. Formalmente:

$$R = \frac{|\text{documentos relevantes recuperados}|}{|\text{documentos relevantes en la colección}|} \tag{2.1.5}$$

Otra medida ampliamente utilizada es la *precisión* (en Inglés *precision*), denotada por la letra P, la cual se define como el número de documentos relevantes sobre el número de documentos recuperados. Formalmente:

$$P = \frac{|\text{documentos relevantes recuperados}|}{|\text{total de documentos recuperados}|} \tag{2.1.6}$$

Dependiendo de la aplicación, una de las medidas anteriores puede ser más importante que la otra. Por ejemplo, cuando se trata de sistemas de RI que actúan sobre el internet, los desarrolladores se preocupan por que el usuario obtenga el mayor número posible de documentos relevantes dentro de los resultados presentados, es decir, prefieren una precisión alta. Sin embargo, cuando un sistema de RI se utiliza para obtener la mayor cantidad de información relevante dentro de una colección para realizar un análisis posterior, lo que se prefiere es un recuerdo alto aunque la precisión sea baja. Para evitar esta situación y tratar de obtener una medida más representativa, se cuenta con medidas que combinan la precisión y el recuerdo.

Una de estas medidas es llamada *medida F* (en Inglés *F-measure*), la cual se define de la siguiente manera:

$$F(P,R) = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{donde} \quad \beta^2 \in [0,\infty) \tag{2.1.7}$$

Podemos verificar de manera sencilla que $\lim_{\beta\to 0} F(P,R) = P$, y que $\lim_{\beta\to\infty} F(P,R) = R$, por lo que valores bajos de β dan más importancia a la precisión, mientras que valores altos de β dan más importancia al recuerdo. La figura

2.4 muestra el comportamiento de esta medida considerando una precisión de 0.8 y un recuerdo de 0.3.

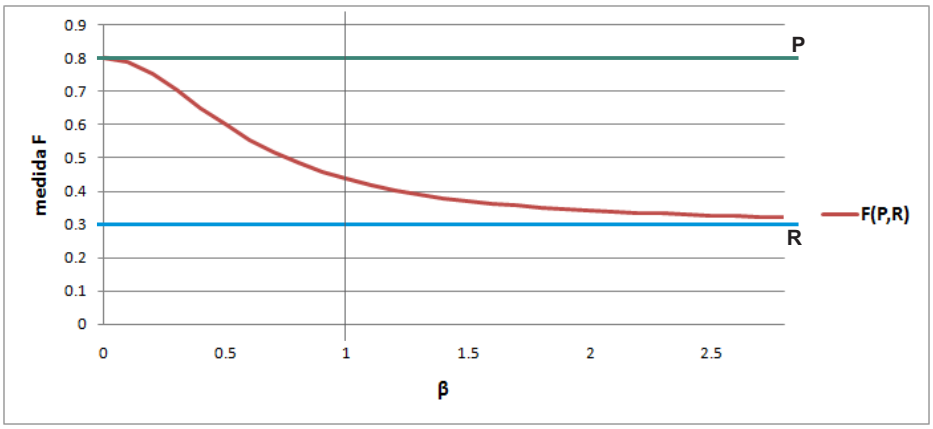

Figura 2.4: Comportamiento de la medida F

La versión más utilizada de esta medida es dando a β un valor de 1, la cual es llamada *medida F balanceada*. Con lo anterior, la fórmula 2.1.7 se simplifica a:

$$F_{\beta=1} = \frac{2PR}{P+R} \tag{2.1.8}$$

En la figura 2.4 podemos observar que la medida F balanceada es más rigurosa que el promedio para los valores definidos de P y R. En general, cuando más parecidos sean los valores de las medidas P y R, la medida F tenderá al valor promedio de las mismas.

Las medidas anteriores no toman en cuenta de manera directa la posición en la que los documentos relevantes se encuentran, lo cual es una característica importante. Una medida que toma tanto el número de documentos relevantes como su posición en la lista es la medida *Precisión a r documentos* (*r-prec*). Esta medida se define como:

$$r\text{-}prec = \frac{1}{r}\sum_{i=1}^{r} rel(i) \tag{2.1.9}$$

donde

$$rel(i) = \begin{cases} 1, & \text{Si } d_i \text{ es un documento relevante;} \\ 0, & \text{en otro caso.} \end{cases} \tag{2.1.10}$$

y r es el número de documentos relevantes en la colección para la petición evaluada.

Otra medida con estas cualidades es la medida AP (*Average Precision*). Esta

considera la precisión de diferentes subconjuntos de la lista, tomando como puntos para cada subconjunto la posición de cada documento relevante encontrado, y hace un promedio de dichas precisiones. Formalmente:

$$AP = \frac{\sum_{r=1}^{n} \text{r-prec} \times rel(d_r)}{m} \tag{2.1.11}$$

donde n es el número de documentos recuperados, *r-prec* es la precisión a r documentos (definida en 2.1.9), $rel(d_r)$ definida como en 2.1.10, y m es el número de documentos relevantes existentes en la colección completa para la petición evaluada. Ésta es la medida más utilizada actualmente para medir la eficacia de los sistemas de RI para cada petición.

Para medir la eficacia global de un sistema de RI en un conjunto de prueba, la medida más utilizada es el *MAP* (*Mean Average Precision*), el cual es el promedio aritmético del *AP* obtenido para cada petición. Formalmente:

$$MAP = \frac{\sum_{i=1}^{p} AP_i}{p} \tag{2.1.12}$$

Donde p es el número de peticiones en el conjunto de prueba. Más información acerca de estas medidas, así como medidas adicionales pueden consultarse en [11].

Tanto el *AP* como el *MAP* son las medidas de evaluación utilizadas para medir la eficacia del método presentado en esta investigación, debido a su utilidad comprobada en diferentes artículos y foros de evaluación.

2.2. Fusión de datos

Debido a la variedad de formas en las que un usuario puede realizar una consulta, a las diferentes colecciones donde puede actuar un sistema de RI y a la variedad de la información digitalizada (documentos, imágenes y audio), no existe un sistema de RI que ofrezca buenos resultados en todos estos escenarios. Por esta razón el desarrollo de sistemas de RI ha ido en aumento, haciendo que en nuestros días exista una gran variedad de ellos. Diferentes enfoques se han propuesto para el desarrollo de sistemas de RI, por ejemplo los enfoques basados en PLN que intentan hacer un "entendimiento" del contenido de los documentos y de la petición para realizar la comparación que identifique a los documentos relevantes. Los enfoques estadísticos utilizan características como frecuencias y posición de las palabras para determinar

su importancia en el documento.

Dependiendo de la petición y la colección de elementos, algunos enfoques pueden desempeñarse mejor que otros, sin embargo siempre existirá algún caso en el que suceda lo contrario.

Puede pensarse que en un futuro se contará con un sistema de RI que logre tratar cualquier petición en cualquier colección con una eficacia aceptable, sin embargo una solución alternativa se ha encontrado en una tarea llamada *Fusión de Datos*.

2.2.1. Definición de la tarea

Hsu y Taksa [16] definen la tarea de Fusión de Datos de la siguiente manera:

> Fusión de Datos es un proceso (adquisición, diseño e interpretación) de combinación de información obtenida por múltiples agentes (fuentes, esquemas, sensores o sistemas) en una simple representación (o resultado).

La Fusión de Datos ha sido utilizada en campos como el reconocimiento de patrones y aprendizaje automático, tomando como información a fusionar las diferentes características extraídas; en detección de señales, rastreo de blancos, procesamiento de imágenes, vigilancia y aplicaciones militares, donde los datos arrojados por sensores son la información que se fusiona [16].

En el área de Recuperación de Información, el concepto de Fusión de Datos ha sido utilizado para estudiar la combinación de múltiples resultados, obtenidos de diferentes formulaciones de la petición o de diferentes sistemas. Existen diferentes esquemas que pueden ofrecer diferentes resultados de una misma petición [16]. Las figuras 2.5, 2.6 y 2.7 muestran los posibles escenarios. En éstas figuras, **p** representa una petición, **R** es un proceso de reducción a la representación de la colección, $\mathbf{p}_r$ es la petición reducida, **SRI** es un sistema de RI, **r** es un resultado de recuperación, **S** es un proceso de selección del mejor resultado de recuperación y $\mathbf{r}_f$ es el resultado final.

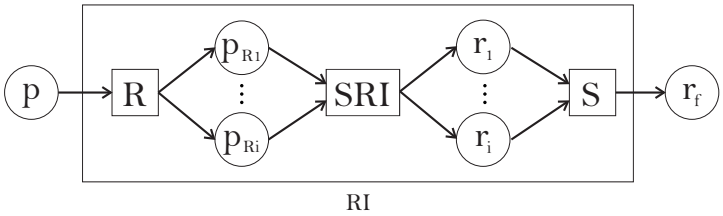

Figura 2.5: Proceso de Recuperación de Información: multiples formulaciones y un sólo sistema

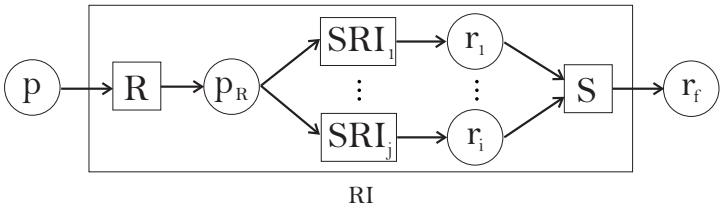

Figura 2.6: Proceso de Recuperación de Información: una formulación y múltiples sistemas

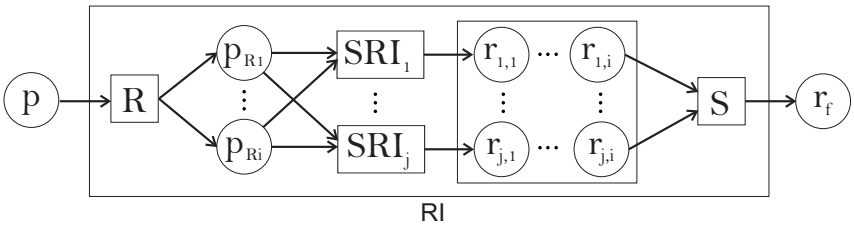

Figura 2.7: Proceso de Recuperación de Información: múltiples formulaciones y múltiples sistemas

Los esquemas más utilizados se basan en múltiples formulaciones de la petición procesadas por un único sistema de RI (figura 2.5) y la utilización de múltiples sistemas de RI para procesar una misma formulación de la petición (figura 2.6). Otra posibilidad, aunque no muy investigada actualmente, es una combinación de las dos modificaciones anteriores, es decir, tener múltiples formulaciones de la petición y procesarlas con múltiples sistemas de RI (figura 2.7). En los escenarios anteriores, el resultado ofrecido al usuario es del SRI que muestra una mayor eficacia en una colección de entrenamiento (bloque S en los diagramas). Una alternativa es utilizar Fusión de Datos, ya que estos escenarios son, de manera natural, el campo de acción de esta tarea.

Pero ¿Cuál es el objetivo de la Fusión de Datos? Teniendo en mente que para una misma petición diferentes sistemas pueden tener una eficacia distinta, el objetivo de la Fusión de Datos es mejorar los resultados de recuperación de los SRIs.

En un escenario real, donde se cuenta sólo con las listas de resultados sin ninguna información adicional acerca de su eficacia para satisfacer la necesidad de información, la Fusión de Datos intenta enfatizar las fortalezas de las listas y disminuir sus debilidades. Por tanto, lo que se espera de la Fusión de Datos es que su resultado final sea mejor que el de la mayoría los SRIs, preferentemente que el de todos, y que en ningún caso empeore los resultados de éstos.

En la práctica, el objetivo de la Fusión de Datos no siempre se logra, por lo que diferentes estudios se han realizado, los cuales tratan de determinar cuándo la

Fusión de Datos es aplicable a un conjunto de listas de resultados. Vogt y Cottrell [9] investigaron la factibilidad de predecir la eficacia de la fusión de pares de listas; Ng y Kantor [37] consideraron el problema de determinar cuándo la fusión de dos listas supera la eficacia de ambas; Wu y Mclean [55] extendieron la investigación anterior considerando tres conjuntos de datos, tres métodos de fusión, y fusiones de 3 a 10 listas de resultados. Un modelo que aplica una función de pesado distinta para cada petición en lugar de utilizar una misma para todo el conjunto, fue propuesto por Diamond y Liddy [53]. Sin embargo, el problema de determinar cuántas y qué listas deben fusionarse permanece abierto.

Diferentes métodos de fusión se han propuesto, los cuales han sido utilizados en los análisis mencionados. A continuación se presenta una clasificación de estos métodos y se describen los que fueron utilizados en esta investigación.

2.2.2. Métodos de fusión

Dentro de la tarea de Fusión de Datos se han propuesto diferentes métodos de fusión que se basan en diferentes características de los elementos dentro de las listas de resultados. Vogt y Cottrell [9, 12] presentan tres efectos que suelen darse en las listas de resultados:

- **Efecto skimming**. En este efecto diferentes enfoques de recuperación que representan a sus elementos de manera distinta, pueden recuperar diferentes elementos relevantes. Por tanto, un método de fusión que tome los elementos mejor posicionados de cada uno de los enfoques de recuperación, tenderá a colocar a los elementos no relevantes en posiciones posteriores en la lista.

- **Efecto coro**. Este efecto ocurre cuando varios enfoques de recuperación sugieren que un elemento es relevante para una petición. Esto tiende a ser una evidencia más fuerte para la relevancia del elemento, que la sugerencia de un sólo enfoque de recuperación. Un método de fusión que tome los elementos más redundantes tendrá más probabilidades de ofrecer mejores resultados de recuperación.

- **Efecto caballo negro**. En este efecto un enfoque de recuperación puede producir estimaciones de relevancia muy altos o muy bajos para algunos elementos, en comparación con otros enfoques de recuperación. Este efecto suele afectar

seriamente los resultados de los métodos de fusión cuando se incluyen listas con resultados pobres.

Para tener un panorama más sencillo de estos efectos, podemos relacionar el efecto skimming con la *complementariedad*, la cual se presenta cuando, para una misma petición, un sistema de RI recupera elementos relevantes que otro no logró recuperar y viceversa. El segundo efecto, coro, se relaciona con la *redundancia* la cual se presenta cuando diferentes sistemas de RI recuperan elementos en común para una petición. El tercer efecto, caballo negro, podemos relacionarlo con el valor de relevancia que los sistemas de RI dan a cada elemento de acuerdo a la medida de similitud que utilicen. Los métodos de fusión utilizan una o más de las características anteriores. El funcionamiento de los métodos de fusión sigue los siguientes pasos generales:

1. Normalizar los pesos de los elementos en las listas.

2. Aplicar una fórmula para combinar los pesos de los elementos comunes en las listas, y así obtener nuevos pesos que serán utilizados en la lista resultante para hacer un reordenamiento.

3. Reordenar los elementos y construir la lista final.

Lo que hace diferentes a los métodos de fusión es la forma en la que normalizan los pesos y la fórmula que utilizan para combinarlos. A continuación se presenta una clasificación de los métodos de fusión y, en el caso de aquellos utilizados en este trabajo, el detalle de su funcionamiento.

Métodos de fusión lineales

Estos métodos utilizan los valores de relevancia dados por los sistemas de RI a los elementos en la listas, para utilizarlos en una función lineal. Tres son los métodos básicos de esta clase de métodos: MaxRSV, MinRSV y SumRSV [46] (RSV son las siglas de *Retrieval Score Value* o también puede encontrarse como *Raw Score Value*). El funcionamiento de estos métodos se basa en la redundancia de los elementos en las listas, sin embargo, el valor de similitud dado por los SRIs juega un papel fundamental ya que determina la posición de los elementos en la lista final. La tabla 2.1 muestra las operaciones de estos tres métodos:

Método	Función
MinRSV	$min\{s_1, s_2, ..., s_n\}$
MaxRSV	$max\{s_1, s_2, ..., s_n\}$
SumRSV	$\sum_{i=1}^{n} s_i$

Tabla 2.1: Métodos lineales básicos

En la tabla anterior, $s_1, s_2, \ldots, s_n$ representan los diferentes *scores* para un elemento recuperado en las diferentes listas de resultados; n es el número de listas que se fusionan, por lo que pueden existir mínimo un RSV para algún elemento y a lo más n. En caso de que en alguna lista no exista un elemento, simplemente se descarta esa lista en las funciones Max y Min. En el caso de la función Sum, si un elemento no existe en una lista, se utiliza el valor 0 para esa lista. En esta investigación, el método utilizado de esta clase es MaxRSV, ya que otorga una importancia mayor al ordenamiento dado por los SRIs. A continuación se muestran los detalles de la implementación realizada.

MaxRSV

Sea $L = \{l_1, ..., l_i, ..., l_m\}$ un conjunto de listas de resultados ordenados de la forma $l_i =< d_1, ..., d_j, ..., d_n >$ donde d_j representa a un documento recuperado. Sea $D = \{l_1 \cup ... \cup l_i \cup ... \cup l_m\}$ el conjunto unión de los elementos en las listas de L. El nuevo peso para los elementos en D se define como:

$$maxRSV(d_j) = max\{s(d_j, l_i), \forall l_i \in L\} \tag{2.2.1}$$

donde $s(d_j, l_i)$ es el peso normalizado del elemento d_j en la lista l_i, o cero si d_j no se encuentra en l_i. En nuestra implementación, la normalización de los pesos de cada lista se realizó dividiendo todos los pesos entre el peso más alto.

Detalles sobre métodos lineales adicionales pueden consultarse en [6, 10, 12, 34].

Métodos de fusión posicionales

Estos métodos descartan completamente los RSVs de los elementos en las listas y calculan nuevos pesos iniciales a partir de diferentes estrategias. Los métodos más representativos de esta clase se muestran en la tabla 2.2:

Los métodos CombMIN, CombMAX y CombSUM son iguales a sus contrapartes

Método	Fórmula
CombMAX	$max\{s_1, s_2, ..., s_n\}$
CombMIN	$min\{s_1, s_2, ..., s_n\}$
CombSUM	$\sum_{i=1}^{n} s_i$
CombANZ	$\frac{\sum_{i=1}^{n} s_i}{\text{número de pesos distintos de cero}}$
CombMNZ	$(\sum_{i=1}^{n} s_i) *$ número de pesos distintos de cero
CombMED	$mediana\{s_1, s_2, ..., s_n\}$

Tabla 2.2: Métodos posicionales

lineales (minRSV, maxRSV y sumRSV), la diferencia son los pesos utilizados en el proceso (la familia de los RSVs considera los valores de similitud dados por los SRIs, mientras que la familia de los Comb utiliza las posiciones de los elementos).

CombMIN "desconfía" del criterio de los SRIs al asignar una similitud alta a los elementos, por lo que si un elemento aparece en más de una lista, elige al de la posición más baja. Por otro lado, CombMAX confía en los SRIs por lo que elige al elemento mejor posicionado. Estos dos métodos son contradictorios ya que CombMIN promueve el tipo de error que CombMAX trata de evitar y viceversa. El método CombMED es un enfoque sencillo que evita ambos escenarios al tomar el valor medio de los pesos. Los métodos anteriores sólo seleccionan un peso, ignorando la información que otorgan los pesos descartados. El método CombSUM utiliza la información de todos los pesos al sumar sus valores. El método CombANZ realiza el promedio de los pesos distintos de cero, lo cual ignora el efecto de un sistema al no recuperar documentos para una petición. Por último, el método CombMNZ otorga pesos más altos a aquellos elementos presentes en varias listas.

El método CombMNZ ha demostrado ser el mejor de esta clase [35], además de ser considerado el método base para la tarea de Fusión de Datos, razón por la cual fue elegido para realizar los experimentos en esta investigación. Los detalles de la implementación se muestran a continuación.

CombMNZ

Sea $L = \{l_1, ..., l_i, ..., l_m\}$ un conjunto de listas de resultados ordenados de la forma $l_i =< d_1, ..., d_j, ..., d_n >$ donde d_j representa a un documento recuperado. Sea $D = \{l_1 \cup ... \cup l_i \cup ... \cup l_m\}$ el conjunto unión de los elementos en las listas de L. El nuevo peso para los elementos en D se define como:

$$CombMNZ(d_j) = \left(\sum_{i=1}^{|L|} e(d_j, l_i)\right)\left(\sum_{i=1}^{|L|} s(d_j, l_i)\right) \tag{2.2.2}$$

donde

$$e(d_j, l_i) = \begin{cases} 1, & \text{si } d_j \in l_i \\ 0, & \text{en caso contrario} \end{cases} \tag{2.2.3}$$

$$s(d_j, l_i) = |l_i| - r(d_j, l_i) + 1 \tag{2.2.4}$$

y $r(d_j, l_i)$ es la posición del elemento d_j en la lista l_i. En nuestra implementación, el RSV de cada elemento se sustituyó por su posición en la lista, y es este valor el utilizado en la función r.

Más detalles sobre este método así como información adicional de otros métodos posicionales pueden consultarse en [21, 34, 35, 43].

Métodos de fusión basados en la teoría de elección social

Esta clase de métodos se basan en una analogía de la teoría de elección donde los documentos son considerados los candidatos y cada lista de resultados un votante. Dos son los métodos representativos de esta clase, los cuales son descritos a continuación:

- **Conteo Borda**. Es un algoritmo que otorga un peso a cada candidato de acuerdo a la posición dada por los votantes a favor del candidato. En este algoritmo, para cada votante, el candidato en la primera posición recibe n puntos (donde n son el número de candidatos en la elección), el segundo recibe $n-1$ y así sucesivamente. El candidato con más puntos gana, pero en el caso de RI, los documentos (o candidatos) son reordenados de acuerdo al número de puntos obtenidos.

- **Votación Condorcet**. Este algoritmo especifica que el ganador de la elección es el candidato, o candidatos, que vence o empata con cualquier otro candidato en una comparación uno a uno. El resultado de este algoritmo es una tabla donde se especifican el número de comparaciones ganadas, empatadas y perdidas. En RI, esta tabla es utilizada para hacer el ordenamiento de los documentos (o candidatos), tomando como primer criterio el número de comparaciones ganadas,

como segundo criterio el número de comparaciones empatadas y como último criterio el número de comparaciones perdidas.

Existen diferentes implementaciones de estos algoritmos aplicados a la Recuperación de información, en particular una variante del conteo Borda, llamado *conteo Borda difuso* (en Inglés *fuzzy Borda count*), fue utilizado en esta investigación debido a que sus resultados son equivalentes y algunas veces superiores a los del método CombMNZ. Los detalles del método *conteo Borda difuso* se muestran a continuación.

Conteo Borda difuso (Fuzzy Borda count)

En este método, consideraremos a $L = \{l_1, ..., l_i, ..., l_m\}$ como el conjunto de m expertos (votantes) los cuales dan su preferencia (voto) a sus alternativas (candidatos) $l_i =< d_{i1}, ..., d_{ij}, ..., d_{in} >$ representadas por $p_i =< w_{i1}, ..., w_{ij}, ..., w_{in} >$. Utilizando las preferencias dadas por cada experto se genera una *matriz de intensidad de preferencias* de la forma

$$\begin{pmatrix} r^i_{11} & r^i_{12} & \cdots & r^i_{1k} & \cdots & r^i_{1n} \\ r^i_{21} & r^i_{22} & \cdots & r^i_{2k} & \cdots & r^i_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ r^i_{j1} & r^i_{j2} & \cdots & r^i_{jk} & \cdots & r^i_{jn} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ r^i_{n1} & r^i_{n2} & \cdots & r^i_{nk} & \cdots & r^i_{nn} \end{pmatrix} \tag{2.2.5}$$

donde

$$r^i_{jk} = \frac{w_{ij}}{w_{ij} + w_{ik}} \tag{2.2.6}$$

En la matriz de intensidad de preferencias, cada r^i_{jk} corresponde al grado de confianza con el cual cada experto i prefiere a la alternativa d_j sobre la alternativa d_k. El valor final de confianza asignado por el experto i para cada alternativa d_j es la suma en filas de los valores mayores a 0.5 en la matriz de intensidad de preferencias. Formalmente:

$$r_i(d_j) = \sum_{k=1, r^i_{jk} > 0{,}5}^{n} r^i_{jk} \tag{2.2.7}$$

El valor del conteo Borda difuso final asignado a cada candidato, es la suma de los valores asignados por cada experto. Formalmente:

$$r(d_j) = \sum_{i=1}^{m} r_i(d_j) \tag{2.2.8}$$

Por último, para formar la lista final, los candidatos son ordenados de acuerdo al su valor $r(d_j)$.

Información adicional acerca de este tipo de métodos puede consultarse en [34, 36, 40, 41].

Métodos de fusión probabilísticos

Este tipo de métodos suponen que la eficacia de los sistemas de RI, en un número de peticiones de prueba, es un indicativo de su eficacia futura. Durante el proceso de entrenamiento se calculan las probabilidades de relevancia mediante el siguiente proceso:

1. Cada lista l_i de resultados es dividida en un número de segmentos.
2. Se calcula la probabilidad condicional de la relevancia (R) de cada documento d_j dependiendo en que segmento k ocurra, esto es $prob(R(d_j)|d_j, k, l_i)$.

Para peticiones subsecuentes, el peso para cada documento d_j normalmente está dado por $\sum_{i=1}^{m} \frac{prob(R(d_j)|d_j,k,l_i)}{k}$, esto es, la suma de la probabilidad de que el documento d_j ocurra en el segmento k en todas las listas de resultados.

Este tipo de métodos dependen de un entrenamiento previo, y debido a que la presente investigación tiene como objetivo el desarrollo de un método que sea independiente del funcionamiento interno de los sistemas de RI, los métodos probabilísticos no fueron utilizados en nuestros experimentos. Más información acerca de estos métodos puede encontrarse en [14, 28, 34, 40]

2.2.3. Evaluación

Debido a que el resultado de todo método de fusión es una lista de resultados, la forma de evaluación de la eficacia de estos métodos para satisfacer una necesidad de información, es la misma que la utilizada para de los sistemas de RI (ver sección 2.1.4).

Capítulo 3

Trabajo Relacionado

En el capítulo 1 se describió el problema que se trata en esta investigación, el cual puede resumirse en la selección a priori de las listas que al fusionarse, ofrezcan mejores resultados de recuperación que la fusión de todas las listas disponibles. Este problema involucra algún tipo de predicción de la eficacia de las listas individuales para poder descartar aquellas que pueden afectar a la Fusión de Datos debido resultados de recuperación pobres. Este capítulo muestra diferentes formas en las que se ha atacado este problema, las cuales tienen una relación directa o indirecta con la problemática principal de esta tesis.

3.1. Distintas posibilidades de solución

En el estado del arte podemos encontrar distintos enfoques que tratan de determinar la eficacia de un sistema de RI. A continuación se describen cuatro de ellos:

- **Ordenamiento de Sistemas de RI.** El propósito de esta tarea es realizar un ordenamiento de un conjunto de sistemas de RI de acuerdo a su eficacia, el cual se obtiene con diferentes estrategias que evitan la dependencia de los juicios de relevancia[1] [1, 47, 50].

- **Selección de la mejor lista.** En esta tarea se obtienen listas de resultados para una petición, mediante diferentes procesos se califican y, por último, se selecciona la mejor para dicha petición[13, 27].

[1]Los juicios de relevancia son una lista de los elementos relevantes existentes para las peticiones en la colección de búsqueda. Se construyen de manera manual y sólo están disponibles para colecciones de evaluación (ver sección 2.1.4).

- **Análisis de la eficacia de la fusión.** En esta tarea se generan las fusiones posibles dado un conjunto de listas iniciales, se obtienen características de las listas iniciales y de las fusiones, y, mediante diferentes estrategias, se trata de determinar la efectividad que tendrá la fusión, y si esta eficacia puede superar al de las listas iniciales [9, 37, 55].

- **Selección de listas para la fusión.** En esta tarea diferentes listas de resultados para una petición son evaluadas para determinar su utilidad en el proceso de fusión. Las más aptas son utilizadas para generar la lista de resultados final [45, 53].

De las tareas anteriores, las dos últimas tienen relación directa con la problemática principal de esta investigación. Las dos primeras presentan un objetivo distinto, sin embargo, sus estrategias de estimación de la eficacia son un punto común con nuestra problemática. A continuación se describen los trabajos del estado del arte con relación a las tareas descritas anteriormente, y se discute acerca de la relación y diferencias que guardan con el método de Fusión Dinámica de Resultados de RI propuesto en esta tesis.

3.2. Ordenamiento de sistemas de RI

El objetivo principal de esta tarea es poder decidir cuándo un sistema de RI es mejor que otro, dado un conjunto de datos de prueba, sin utilizar los juicios de relevancia para las peticiones del conjunto, con el propósito de hacer un ordenamiento automático de dichos sistemas.

Wu y Crestani[50] realizan la evaluación de sistemas de RI mediante una medida llamada *reference count*. Dado un conjunto de m listas de resultados obtenidos mediante diferentes esquemas de recuperación (sistemas o reformulaciones), cada lista de resultados es calificada de acuerdo a las ocurrencias de sus primeros n documentos en las listas restantes (a dichos documentos se les llama *referencias*), esto es, para la lista l_1, $S_{l_1} = \sum_{j=1}^{n} o(d_{i,j})$, donde n es el número de elementos considerados de la lista l_1 y $o(d_{i,j})$ es la suma de las ocurrencias del elemento j en las listas restantes, con $1 \leq i \leq m$ e $i \neq 1$. Cuatro variaciones a esta medida básica se realizaron para dar mayor peso a los elementos redundantes en posiciones superiores, para considerar una distribución hipotética de los elementos con alta probabilidad de ser relevantes en la

lista de resultados, y para tomar en cuenta los pesos ofrecidos por los esquemas de recuperación. Estas medidas fueron probadas con 5 colecciones del TREC (TREC 3 (40 sistemas), TREC 5 (61 sistemas), TREC 6 (74 sistemas), TREC 7 (103 sistemas) y TREC 2001 (97 sistemas)). El ordenamiento "manual" (o *ranking*) de los sistemas se realizó tomando en cuenta tres medidas de efectividad: *average precision*, *r-prec* y *precision a 100 documentos*. La evaluación del ordenamiento resultante al aplicar sus medidas de *reference count* se realizó mediante el coeficiente promedio de correlación de ordenamiento Spearman (mean Spearman rank correlation) entre el ordenamiento manual y los resultantes de sus medidas. Para el caso de *average precision*, su mejor resultado se obtuvo para el conjunto TREC 3, obteniendo un coeficiente de correlación de 0.586. Para el caso de *r-prec* su mejor resultado fue para el mismo conjunto con un coeficiente de correlación de 0.620. Por último, para la *precision a 100* se obtuvo un coeficiente de 0.641 para el mismo conjunto. El hecho de que sus mejores resultados se obtengan en la tercera medida se debe a que su enfoque se ajusta directamente a esta medida; *r-rpec* es una medida más difícil, ya que no se conoce el número de documentos relevantes para cada petición y su enfoque utiliza un número fijo de elementos para la evaluación; en el caso de *average precision* las posiciones de los elementos relevantes son fundamentales, y *reference count* no es suficientemente exacto localizando las posiciones de los elementos relevantes.

Nuray y Can [47] presentan un enfoque basado en fusión de datos para ordenar sistemas de RI de acuerdo a su eficacia. Debido a que no pueden hacerse uso de juicios de relevancia, los cuales contienen a los elementos relevantes para cada tópico, utilizan Fusión de Datos para obtener elementos llamados "seudo-relevantes". La medida *MAP* de las listas de resultados se obtienen sustituyendo a los juicios de relevancia por los elementos seudo-relevantes, y con base en el resultado de esta medida se ordenan los SRIs. El proceso es elegir a las listas para ser fusionadas, elegir un porcentaje de los documentos mejor ordenados, y por último obtener la medida de evaluación. Tres enfoques se utilizaron para la selección de las listas a fusionar: *Best*, la cual incluye un porcentaje de los mejores sistemas disponibles (calificados a priori mediante los juicios de relevancia); *Normal*, la cual incluye a todos los sistemas disponibles; y *Bias*, la cual incluye un porcentaje de los sistemas que obtienen elementos diferentes a la mayoría de todos los sistemas. Los conjuntos utilizados para realizar los experimentos fueron TREC 3, TREC 5, TREC 6 y TREC 7. Los métodos de fusión *Rango Recíproco*, *Conteo Borda* y *Votación Condorcet* fueron probados para la obtención

de los elementos pseudo-relevantes. Los resultados obtenidos muestran que no hay una combinación de método de selección de listas y de método de fusión que obtenga siempre el mejor coeficiente de correlación de Spearman para todos los conjuntos, si embargo, en promedio, el método votación Condorcet mediante una selección tipo Bias de las listas a fusionar obtienen un 0.635, siendo éste el mejor resultado. Este trabajo concluye que determinar un número adecuado de elementos seudo-relevantes, obtenidos de la fusión de listas adecuadas, mejora el "ranking" final de los sistemas de RI según su eficacia medida en MAP.

Spoerri [1] demuestra que el traslape de los documentos entre las listas generadas por los sistemas de recuperación, es útil para ordenar a los sistemas de acuerdo al número de elementos que un sistema de RI tiene en común con los restantes. Las medida que utilizan para ordenar los resultados es el promedio del porcentaje de documentos que un sistema recupera por sí solo (es decir, que no recuperan los demás sistemas), así como el resultado de este porcentaje menos el porcentaje de elementos que recuperan todos los sistemas. Estos porcentajes son llamados *Single %* y *Single %-AllFive %*. Los porcentajes son obtenidos de grupos de 5 sistemas creado de manera aleatoria. Resultados para los conjuntos TREC 3 (18 sistemas), TREC 6 (24 sistemas), TREC 7 (25 sistemas) y TREC 8 (35 sistemas) fueron utilizados. Sus resultados muestran que los 50 documentos mejor posicionados de cada lista son suficientes para inferir la calidad del sistema. El mejor resultado es un coeficiente de correlación de 0.95 entre el ordenamiento oficial utilizando el MAP de los sistemas del TREC 8 y el ordenamiento conseguido por su método basado en traslape. Como conclusión de este estudio, se muestra que el porcentaje promedio de los elementos de un sistema que sólo son recuperados por él, está fuerte y negativamente correlacionado con su eficacia, tomando ya sea el MAP o p_{1000} (precisión a 1000 documentos recuperados) como medidas de evaluación.

3.3. Selección del mejor resultado de RI

Si bien esta tarea también califica la eficacia de las listas de resultados, el objetivo es distinto, ya que no pretende realizar un ordenamiento de los sistemas sino seleccionar, para cada petición, un solo resultado y ofrecerlo al usuario.

A este respecto, Hubert y Mothe [27] proponen un método basado en la retroalimentación de relevancia para seleccionar al mejor resultado de recuperación para una petición dada. Su método se basa en tomar los primeros n elementos de cada lista de recuperación y proporcionarlos al usuario para que sea él quien determine su relevancia. De esta manera, el sistema con la mayor relevancia es utilizado para tratar dicha petición. En sus experimentos, ellos evitan la necesidad de usuarios reales al sustituir el paso de determinación de relevancia por parte del usuario, por una determinación de los elementos relevantes utilizando los juicios de relevancia. Con lo anterior, el criterio para decidir qué sistema se utilizará para la petición es el valor máximo de *precisión a 5* documentos (P@5). Este método fue probado con los conjuntos TREC 3, TREC 5, TREC 6 y TREC 7, de los cuales se realizaron pruebas considerando 2 y 5 sistemas. Sus resultados muestran una mejora promedio de 9.4 % considerando dos sistemas, y de 9.9 % considerando 5 sistemas.

Por otro lado Kompaoré et al. [13] proponen un enfoque basado en agrupamiento de peticiones para detectar al mejor sistema. Su idea se basa en la hipótesis de que algunos sistemas se desempeñan mejor en algún tipo de peticiones que otros, por lo que agrupar un conjunto de peticiones de entrenamiento de acuerdo a características lingüísticas, e identificar al sistema que mejor se desempeñe en cada grupo, permitirá identificar al sistema a utilizar al tratar las peticiones de un conjunto de prueba. Se utilizaron 13 características lingüísticas de las peticiones: longitud promedio de términos, promedio de morfemas, número de sufijos frecuentes, número de nombres propios, acrónimos, números, palabras desconocidas, conjunciones, preposiciones, pronombres, profundidad sintáctica, tamaño de la liga sintáctica en palabras, y el número de synsets de las palabras en WordNet. El método de agrupamiento utilizado fue AHC (Agglomerative Hierarchical Clustering) con la distancia Euclidiana como medida de distancia entre individuos (peticiones) y el criterio Ward como medida de distancia entre grupos. Se utilizó el Análisis de Componentes Principales como método para reducir la dimensionalidad de los individuos. Para realizar los experimentos se consideraron 200 peticiones extraídas de las colecciones TREC 3, TREC 5, TREC 6 y TREC 7. Se contó con 40, 80, 79 y 103 sistemas para cada colección, respectivamente. El conjunto de entrenamiento constó de 180 peticiones y el de prueba de las 20 restantes. Estos conjuntos se seleccionaron de manera aleatoria en 10 iteraciones. Al aplicar su método al conjunto de prueba, su método logra mejoras mayormente en las medidas de precisión a 5 y 10 documentos, que van de 3.72 % a

5.97 %, y de 1.48 % a 6.73 % respectivamente.

3.4. Análisis de la eficacia de la fusión

Esta tarea tiene el objetivo de conocer la efectividad que tendrá la Fusión de Datos aplicada a un conjunto de listas de resultados, y así poder determinar si dicha efectividad puede o no superar al de la mayoría, o preferentemente a la de todas las listas iniciales del conjunto.

Vogt y Cottrell [9] presentan un método basado en regresión lineal para predecir la eficacia, medida en *Average Precision*, de la fusion de pares de listas. En sus experimentos utiliza 20 tópicos de la tarea adhoc del TREC5. 61 listas de resultados de 1000 elementos para cada tópico fueron consideradas. Con estos datos se contó con 36600 posibles pares de listas para realizar la predicción. El método de fusión utilizado por Vogt fue RSV (Retrieval Status Values). Este método reordena los documentos de las listas a fusionar de acuerdo a un peso p definido como:

$$p(w_1, w_2, d, q) = w_1 \times p_1(d, q) + w_2 \times p_2(d, q) \tag{3.4.1}$$

donde w_1, w_2 son pesos dados a los sistemas, calculados previamente, según su efectividad para satisfacer las necesidades de información; $p_1(d, q)$ y $p_2(d, q)$ son los valores dados por los sistemas al documento d al procesar la petición q.

En este trabajo se consideran 15 características utilizadas con regresión lineal como variables de predicción, y el AP de la mejor fusión de los pares de listas como objetivo. La medida de evaluación es r^2, la cual toma valores entre 0 y 1 e indica qué tan bien la regresión lineal aproxima a los puntos reales.

Las características consideradas son: AP de cada una de las listas, similitud de "rankings" entre la lista de resultados de cada sistema y los juicios de relevancia, similitud de "rankings" entre el par de listas, el número de elementos en la intersección de las listas, el coeficiente de correlación de los scores de los documentos en la intersección de las dos listas, el número de documentos únicos en cada lista, traslape de documentos relevantes, traslape de documentos no relevantes; similitud de "rankings", intersección, y coeficiente de correlación sólo considerando documentos relevantes; y similitud de "rankings" de documentos no relevantes.

Para realizar el entrenamiento se utilizaron 29280 pares de listas (80 % del total).

Al realizar la predicción en el 20 % de pares de listas restante, el mejor resultados es de r^2 =0.94, lo cual indica que las características utilizadas son buenas para predecir el la efectividad de los pares de listas fusionados. Con los experimentos realizados concluyen que este método obtiene buenos resultados cuando: al menos una de las dos listas contiene buenos resultados de recuperación, ambas listas contienen conjuntos similares de documentos relevantes, las listas contienen diferentes conjuntos de documentos no relevantes, los scores de las listas tienen una distribución similar pero contienen un ordenamiento diferente de los documentos relevantes.

Ng y Kantor [37] presentan un trabajo que trata un problema similar al de Vogt y Cottrell pero con una diferencia fundamental: predecir cuándo el resultado de la fusión de dos listas supera a la mejor de ellas, utilizando la precisión a 100 documentos (p_{100}) como medida base de eficacia. Utiliza dos conjuntos de datos: el del TREC4 routing task como conjunto de entrenamiento, y el del TREC5 routing task como conjunto de prueba. Los elementos utilizados de estos conjuntos fueron 50 tópicos y 26 listas de resultados que generan 16250 posibles fusiones de pares de listas, para TREC4 (entrenamiento); 45 tópicos y 23 listas de resultados, que generan 11385 posibles fusiones de pares de listas para TREC5 (prueba). Como se dijo anteriormente, la medida p_{100} es utilizada como medida de efectividad, por lo que los conjuntos de entrenamiento y prueba se dividen en casos positivos, si la fusión de las dos listas tiene una precisión a 100 documentos mayor que cualquiera de las dos listas que la generan, y negativos en caso contrario. Los casos donde la eficacia es igual son descartados. Con lo anterior, el conjunto TREC4 (entrenamiento) contiene 3623 casos positivos y 9171 casos negativos; para el conjunto TREC5 (prueba) no se ofrecen estos datos.

La efectividad de la fusión de dos listas ($E(S_1 f S_2)$) está dada por la ganancia relativa en p_{100}, definida como:

$$E(S_1 f S_2) = \frac{p_{100}(S_1 f S_2) - max\{p_{100}(S_1), p_{100}(S_2)\}}{max\{p_{100}(S_1), p_{100}(S_2)\}} \tag{3.4.2}$$

La medida utilizada para hacer la fusión es la suma normalizada de medidas de relevancia de los documentos en las listas. Los elementos de las listas son ordenados de acuerdo a esta medida. La normalización en cada lista se define como:

$$s(d^j) = \frac{v(d^j) - v(d^{1000})}{v(d^1) - v(d^{1000})} \tag{3.4.3}$$

donde $v(d)$ es el valor de similitud dado por el SRI al elemento d, y d^j es el documento en la posición j de la lista ordenada de resultados.

En este trabajo, en contraste con el de Vogt y Cottrell, sólo se utilizan dos características extraídas de pares de listas: *ratio* ($\mathbf{r}$) definido como $r = \frac{p_l}{p_h}$, donde p_l es la menor precisión a 100 documentos y p_h es la mayor precisión a 100 documentos. La otra características es la medida de *pares fuera de orden* entre las listas de resultados. Esta medida es obtenida contando los pares de elementos donde se cumple que un elemento A esté posicionado antes del elemento B en una lista, mientras que en la otra B esté antes del elemento A. Estas características se utilizaron como variables predictivas de la efectividad de la fusión en un modelo de regresión logística. Se obtuvo un 76 % de detección de los elementos positivos en el conjunto de entrenamiento, y un 69 % en el conjunto de prueba. Considerando que sólo se utilizan dos características, el análisis realizado en este trabajo muestra que es posible determinar cuándo una fusión puede superar a la mejor lista inicial, sin embargo, no ofrece una evaluación final de los resultados aplicando su predicción a un conjunto de peticiones.

Wu y McClean[55] presentan el trabajo más completo en el problema de determinar cuándo una fusión de listas supera a la mejor de las listas que generan la fusión. Este trabajo amplía las expectativas de los trabajos de Vogt y Cottrell, y Ng y Kantor, ya que considera más conjuntos de datos (TREC6 ad hoc track, TREC2001 web track y TREC2004 robust track), fusiones generadas con más de dos listas (3-10 listas), y 3 métodos de fusión (combSUM, combMNZ y roundrobin). La cantidad de tópicos para cada conjunto son 50, 50 y 249, respectivamente. La cantidad de listas de resultados considerada para cada conjunto es 42, 58 y 77, respectivamente. Wu utiliza sólo 5 características extraídas por cada grupo de listas que conforman la fusión: número de listas utilizada en la fusión (num), traslape de elementos en las listas de resultados (o_rate), promedio de los MAP de las listas de resultados (m_av), desviación estándar de los MAP de las listas de resultados (dev) y el MAP de la mejor lista de resultados ($best$). Se utiliza regresión lineal tomando las primeras 4 características como las variables dependientes para predecir el valor de dos variables dependientes: la eficacia de la fusión, y la ganancia de la eficacia de la fusión sobre el promedio de la eficacia de las listas. La quinta características es incluida para predecir la ganancia de la fusión sobre la efectividad de la mejor lista. Wu realiza una variación de los valores de las características, aplicándoles funciones no lineales e incluyéndolas en las regresiones. El conjunto final de características es: num, $ln(num)$, o_rate, $(o\_rate)^2$,

m_av, $\sqrt{m\_av}$, dev y $(dev)^2$. Los valores de r^2, medida que toma valores entre 0 y 1 e indica que tan bien la regresión lineal aproxima a los puntos reales, resultantes de la regresión lineal para cada conjunto de datos, para cada método de fusión, para cada variable dependiente utilizando las características originales y las modificadas, se presentan a continuación.

TREC6	**combSUM**	**combMNZ**	**roundrobin**
Variables originales	0.848	0.852	0.924
Conjunto aumentado	0.914	0.916	0.930
TREC2001	**combSUM**	**combMNZ**	**roundrobin**
Variables originales	0.816	0.807	0.837
Conjunto aumentado	0.872	0.863	0.848
TREC2004	**combSUM**	**combMNZ**	**roundrobin**
Variables originales	0.778	0.791	0.894
Conjunto aumentado	0.860	0.871	0.910

Tabla 3.1: Valores de r^2 al predecir la efectividad de la fusión.

TREC6	**combSUM**	**combMNZ**	**roundrobin**
Variables originales	0.565	0.553	0.404
Conjunto aumentado	0.755	0.688	0.804
TREC2001	**combSUM**	**combMNZ**	**roundrobin**
Variables originales	0.534	0.537	0.395
Conjunto aumentado	0.745	0.683	0.797
TREC2004	**combSUM**	**combMNZ**	**roundrobin**
Variables originales	0.693	0.680	0.467
Conjunto aumentado	0.461	0.443	0.506

Tabla 3.2: Valores de r^2 al predecir la ganancia de la fusión sobre la eficacia promedio.

Los resultados de las tablas 3.1, 3.2 y 3.3, muestran dos cosas interesantes: que las características elegidas son más útiles para predecir la efectividad de la fusión de las listas, y que operar los valores de las características con funciones no lineales causa una mejor predicción en todos los casos.

TREC6	combSUM	combMNZ	roundrobin
Variables originales	0.654	0.694	0.889
Conjunto aumentado	0.820	0.839	0.840
TREC2001	**combSUM**	**combMNZ**	**roundrobin**
Variables originales	0.788	0.791	0.827
Conjunto aumentado	0.864	0.863	0.905
TREC2004	**combSUM**	**combMNZ**	**roundrobin**
Variables originales	0.695	0.727	0.738
Conjunto aumentado	0.811	0.835	0.922

Tabla 3.3: Valores de r^2 al predecir la ganancia de la fusión sobre la efectividad de la mejor lista.

3.5. Selección de los resultados de RI más aptos para la fusión

Esta tarea tiene como objetivo realizar una fusión "inteligente", realizando un análisis previo de las listas que serán fusionadas con el propósito de evitar una baja efectividad de la Fusión de Datos al considerar en la fusión una lista resultados de recuperación muy malos. Aunque este hecho parece ser muy evidente, pocos son los trabajos que han tratado esta tarea.

Diamond y Liddy [53] proponen un modelo de *Fusión Dinámica*. Este concepto se centra en el uso de una función de fusión lineal diferente para cada petición en lugar de utilizar una única función para todas las peticiones. El modelo propuesto considera la suma de los valores de relevancia que ofrece cada sistema de RI a un elemento en común (como método SumRSV descrito en el capítulo 2), pero además introduce un peso para cada valor de relevancia. Este peso es el centro del dinamismo del método, ya que varía de acuerdo a diferentes características de la petición, características de los documentos y características de la correlación entre listas de resultados. En este trabajo se muestra la oportunidad de mejora de este modelo, la cual se realizó de manera manual mostrando un 23.8 % de mejora sobre el mejor sistema, y de un 22.3 %, considerando *precisión a 5* documentos, sobre el esquema tradicional de utilizar la misma función de fusión para todas las peticiones.

Wu y Crestani [49] realizan una comparación de métodos de fusión clásicos, SumRSV y CombMNZ, con variaciones pesadas de los mismos. Ésta es la misma idea propuesta por Diamond y Liddy, sólo que la forma en la que pesan a los sistemas

es mediante la medida *reference count* (presentada en la sección *Ordenamiento de sistemas de RI*). De acuerdo al valor obtenido de esta medida, a cada sistema se le clasifica como *bueno* (*good*), *justo* (*fair*) y *pobre* (*poor*). A cada clase se le otorga un peso determinado empíricamente (1.25 , 1.0 y 0.75 respectivamente). De esta manera realizan fusión de datos con métodos clásicos agregando un componente "dinámico" que permite otorgar un peso distinto a cada sistema para una petición específica. De acuerdo a sus resultados, sus métodos prueban ser tan buenos como los clásicos, e incluso superiores para ciertos conjuntos de listas de resultados.

Por último Gopalan y Batri [45] proponen un método supervisado para seleccionar las mejores m listas de resultados y el mejor método de Fusión de Datos para una colección de documentos determinada. En este trabajo, todas las posibles fusiones de dos y tres listas de resultados fueron generadas para ser utilizadas en un algoritmo genético, el cual utilizó la medida MAP de las fusiones como su función de aptitud. Cuatro funciones de combinación (fusión) fueron utilizadas en sus experimentos, llamadas C-maxmax, C-maxmin, C-minmax y C-minmin, las cuales se basan en las diferencias de los valores de relevancia dados a los documentos por las estrategias de recuperación. El esquema identificado, es decir, las m listas y el método de fusión son utilizadas en todas las peticiones. Sus resultados muestran un 8.4 % de ganancia sobre el mejor esquema de recuperación.

3.6. Discusión

El estado del arte nos muestra diferentes trabajos que tienen que ver con la predicción de la efectividad de sistemas de RI. También nos muestra diferentes enfoques que han tratado de predecir la eficacia que tendrá el resultado de un método de Fusión de Datos. Sin embargo, existe poca investigación en la tarea de seleccionar las listas que deben fusionarse para obtener una efectividad aceptable. Si bien las estrategias de evaluación de sistemas de RI o de selección de la mejor lista pueden ofrecer una pista de cuáles son las mejores listas de resultados, mostraremos que el fusionar las mejores listas no siempre ofrece mejores resultados de recuperación. Por otro lado, los trabajos realizados en la tarea de análisis de la efectividad de la fusión ofrecen formas de saber cuándo una fusión de listas obtendrá mejores resultados, sin embargo todos dependen de los juicios de relevancia e implementan un enfoque supervisado, lo cual hace difícil su utilización dentro del proceso de RI. Por último, de los pocos trabajos

que implementan un dinamismo en la Fusión de Datos, dos de ellos [53, 49], uno supervisado y otro no supervisado, de alguna manera varían el método de fusión a utilizar mediante el pesado de los valores de relevancia. Sin embargo, como se mostrará posteriormente, el fusionar todas las listas disponibles en la mayoría de los casos no es la mejor estrategia. El último trabajo presenta un enfoque interesante, donde selecciona las listas que serán fusionadas y el método de fusión adecuado para éstas, sin embargo es un enfoque supervisado que forzosamente debe re-entrenarse para cada colección en la que se quiera aplicar.

Las observaciones anteriores motivan este trabajo de tesis, el cual es único en su clase por evitar los factores que limitan a los métodos existentes. Mientras que los enfoques actuales en su mayoría son supervisados, nuestro método es completamente no supervisado; al ser no supervisado evita la dependencia de los juicios de relevancia, los cuales no siempre están disponibles para las peticiones en un conjunto de documentos de búsqueda. Más aún, nuestro método no necesita detalles de las características ni funcionamiento interno de los sistemas de RI con los que se recuperan las listas de resultados a fusionar. Otra característica importante es que, mientras la mayoría de los trabajos mostrados debe generar las posibles fusiones de las listas de resultados antes de seleccionar la mejor de ellas para cada petición, nuestro método evita esta situación y sólo fusiona un subconjunto de listas seleccionadas. Además de lo anterior, nuestro método no está limitado a un sólo método de fusión, ya que la selección de las listas a fusionar son un proceso completamente independiente del método de fusión que desee utilizarse. Por último, nuestro método, al cual hemos llamado *Fusión Dinámica de Resultados de RI*, puede ser incluido de manera sencilla al proceso de RI que genere múltiples resultados de recuperación para una misma petición. A continuación se muestran los detalles del método, los conjuntos de datos utilizados, los análisis, experimentos y resultados obtenidos.

Capítulo 4
Hacia la Fusión Dinámica

En el capítulo 1 se introdujo el concepto de *Fusión Dinámica de Resultados de RI* (FDRI) la cual consiste en determinar, para cada petición, las listas adecuadas para realizar Fusión de Datos con el propósito de ofrecer un mejor resultado que la fusión de todas las listas disponibles.

El trabajo relacionado presentado en el capítulo 3 muestra un panorama de cómo se ha abordado actualmente esta problemática, siendo los trabajos más trascendentes enfoques supervisados que más que resolver el problema, muestran un análisis de la factibilidad del mismo.

Esta investigación tiene su motivación en las observaciones y resultados del trabajo relacionado, ya que si bien se ha mostrado que la fusión puede ofrecer un mejor resultado de recuperación [44], los trabajos en el análisis de la fusión muestran que si pudieran seleccionarse de manera automática las listas que más ayudaran al proceso de fusión, los resultados serían aún mejores que la fusión de todas las listas disponibles.

Con lo anterior en mente, nuestro trabajo tiene la intención de ir un paso más allá de lo que se ha logrado hasta ahora, esto es, automatizar el método de Fusión Dinámica de Resultados de RI e incluirlo en el proceso de RI. Además de la automatización, esta investigación se centra en lograr un método que no tenga las limitaciones de los enfoques actuales, específicamente, el enfoque supervisado, dependencias de los juicios de relevancia y del funcionamiento interno de los sistemas de RI utilizados para recuperar las listas de resultados, y la generación a priori de todas las posibles fusiones para determinar las listas adecuadas para la fusión.

Se ha mencionado que una selección de las listas a fusionar puede mejorar la efectividad de la Fusión de Datos, sin embargo surgen varias interrogantes: ¿En qué magnitud puede esta selección mejorar la efectividad de la fusión de todas las listas? ¿Puede esta selección ofrecer mejores resultados de recuperación que cualquiera de

las listas iniciales del conjunto? y, de acuerdo a las respuestas de las anteriores interrogantes, ¿La posible mejora es realmente significativa?

La siguiente sección presenta una descripción de los conjuntos de datos utilizados en nuestros experimentos, así como un análisis de factibilidad que muestra la utilidad que tendría el método FDRI, con lo cual se responden las interrogantes formuladas anteriormente.

4.1. Colecciones

En nuestros experimentos utilizamos tres diferentes colecciones de elementos, dos de textos y una de imágenes anotadas (imágenes con descripciones textuales de su contenido). Estas colecciones fueron obtenidas del foro de evaluación CLEF [1] y son utilizadas para diferentes tareas del mismo. A continuación se ofrece una descripción de estas colecciones:

LA Times 94. Noticias del año 1994 escritas en Inglés americano. Contiene 113005 elementos en un total de 425 Mb.

Glasgow Herald 95. Noticias del año 1995 escritas en Inglés británico. Contiene 56472 elementos en un total de 154 Mb.

IAPR TC-12. Colección fotográfica de imágenes naturales de diferentes partes del mundo. Incluyen fotografías de diferentes deportes y acciones, personas, animales, ciudades, paisajes y muchos otros aspectos de la vida contemporánea. Contiene 20000 imágenes, cada una con su correspondiente descripción textual.

4.2. Conjunto de peticiones

Antes de presentar los conjuntos de tópicos, es importante definir la diferencia entre un *tópico* y una *petición*. El CLEF define un tópico como la representación de una necesidad de información, la cual contiene diferentes elementos. Este es un ejemplo de un tópico del CLEF:

```
 <topic lang=‘‘en’’>
<identifier>251</identifier>
```

[1] Cross Language Evaluation Forum (*www.clef-campaign.org*)

```
<title>Alternative Medicine</title>
<description>Find documents discussing any kind of alternative
or natural medical treatment including specific therapies such as
acupuncture, homeopathy, chiropractics, or others.
</description>
<narrative>Relevant documents will provide general or specific
information on the use of natural or alternative medical treatments
or practices.
</narrative>
</topic>
```

Los elementos presentes en la representación de la necesidad de información (*tópico*) para esta tarea del foro CLEF son un identificador de lenguaje, un identificador de tópico, un título, una descripción y un narrativa. Por otro lado, una *petición* es una necesidad de información escrita en lenguaje cotidiano. Una petición es, usualmente, la entrada de un sistema de RI. La relación que existe entre tópico y petición es que los elementos de título, descripción y narrativa de un tópico, son utilizados para construir la petición, o peticiones, que serán procesadas por el sistema de RI. Muchas veces sólo el título de cada tópico es utilizado como petición, sin embargo, los lineamientos del foro para las diferentes tareas recomiendan utilizar al menos una combinación del título y la descripción para la construcción de la petición.

Los tópicos de cuatro tareas del foro de evaluación CLEF fueron considerados para nuestros experimentos. La descripción de cada tarea se muestra a continuación.

Ad Hoc CLEF 2005. La intención de esta tarea es realizar peticiones y encontrar documentos relevantes para dichas peticiones en un lenguaje determinado.

Geo CLEF 2008. Recuperación de Información Geográfica (GIR, *Geographic Information Retrieval*) trata a la recuperación de información que envuelve algún tipo de contenido espacial. Dado que muchos documentos contienen algún tipo de referencias espaciales, en ciertas peticiones estas referencias pueden ser importantes, por ejemplo "*find me news stories about riots near Dublin City*".

Image CLEF 2008. Esta tarea evalúa la recuperación de imágenes descritas por fragmentos de texto basadas en peticiones en diferentes lenguajes. Tanto los textos como las imágenes son potencialmente explotables para la recuperación.

Robust CLEF 2008. Esta tarea enfatiza la efectividad estable en todos las peticiones en lugar de una alta eficacia global (busca que en todas las peticiones se obtengan niveles mínimos de efectividad).

La tabla 4.1 muestra la cantidad de tópicos y las colecciones de búsqueda utilizadas para cada una de las tareas descritas anteriormente.

Tarea/Conjunto	Nombre corto	Núm. tópicos	Soportados	Colección
Ad Hoc CLEF 2005	adhoc05	50	50	LA Times 94, Glasgow Herald 95
Geo CLEF 2005	geoir08	25	24	LA Times 94, Glasgow Herald 95
Image CLEF 2005	image08	39	39	IAPR TC-12
Robust CLEF 2008	robust08	160	153	LA Times 94, Glasgow Herald 95

Tabla 4.1: Características de los conjuntos de tópicos

La columna *Soportado* de la tabla 4.1 indica cuántos tópicos contienen al menos un elemento relevante en la colección de documentos. Estos tópicos son los que serán considerados posteriormente en los experimentos. Por otro lado, la columna *nombre corto* muestra la forma en la que, por simplicidad, en adelante nos referiremos a cada conjunto.

Un aspecto importante de los tópicos es el soporte que tienen en la colección de búsqueda, esto es, el número de elementos que pueden satisfacer la necesidad de información que representa cada tópico. Estos elementos reciben el nombre de "relevantes". Entre más elementos relevantes tenga un tópico en la colección de búsqueda, entonces este es mejor soportado por la misma. En los foros de evaluación, para cada tarea se proporciona una lista de los elementos relevantes para cada tópico. Esta lista es obtenida de manera manual mediante una revisión exhaustiva de los documentos en la colección. Esta revisión es realizada por personas quienes evalúan la relevancia de los documentos para la necesidad de información. Esta lista de elementos reciben el nombre de "*juicios de relevancia*" (también pueden ser referidos como *documentos relevantes*, *juicios de evaluación*, *gold standard* o *ground truth*). La tabla 4.2 muestra información acerca del número de elementos relevantes para cada conjunto.

De manera general, entre más elementos relevantes para un tópico existan en la colección de búsqueda, existen más probabilidades de satisfacer su necesidad de información. Por lo anterior se considera a los tópicos con pocos elementos relevantes,

Conjunto	Max.	Min.	Promedio	Desv. Est.
adhoc05	229	3	41.26	41.92
geoir08	109	1	31.12	31.69
image08	184	18	61.56	33.73
robust08	229	1	28.28	34.04

Tabla 4.2: Elementos relevantes en las colecciones de búsqueda

más difíciles que aquellos con muchos elementos relevantes. En la tabla 4.2 podemos observar que el conjunto robust08 es el que contiene más tópicos con pocos elementos relevantes, en promedio. Lo anterior es debido a que este conjunto de prueba fue construido con los tópicos en los que los sistemas de RI participantes en las tareas Ad Hoc del CLEF de los años 2003, 2005 y 2006 se desempeñaron más pobremente. Esto hace a este conjunto de prueba particularmente difícil.

4.3. Sistemas de RI

Para cada conjunto de peticiones se utilizaron 5 diferentes listas de resultados, obtenidas con 5 diferentes sistemas o estrategias de recuperación, las cuales se describen a continuación.

- **adhoc05 y robust08.** Para estos dos conjuntos se utilizó un *toolkit* de Recuperación de Información llamado LEMUR, el cual ofrece las funcionalidades necesarias para crear un índice a partir de una colección de documentos y también ofrece diferentes estrategias de recuperación que pueden ser aplicadas al índice creado previamente. Cinco estrategias de recuperación se utilizaron para obtener las listas de resultados utilizadas en los experimentos (dentro de las configuraciones de LEMUR, los valores del parámetro de modelo de recuperación, *retModel*, fueron: *cosine*, *indri*, *kl*, *okapi* y *tfidf*). Más información acerca de este *toolkit* puede consultarse en *http://www.lemurproject.org*.

- **geoir08.** Las listas de resultados para este conjunto fueron generadas con el toolkit LEMUR y un sistema de RI especializado para tratar consultas geográficas. LEMUR se utilizó para obtener resultados de recuperación base para el Sistema de Recuperación de Información Geográfica (SRIG), mediante el modelo de recuperación *indri*, el cuál se aplicó a tres variantes del conjunto

de peticiones. El primer conjunto de peticiones se generó considerando sólo el título del tópico (T), el segundo considerando el título y la descripción (TD), y el tercero considerando el título, la descripción y la narrativa (TDN). Las listas de resultados obtenidas fueron reordenadas utilizando el SRIG para mejorar los resultados de recuperación. Mas detalles de este sistema pueden consultarse en [20]. Al final,las listas consideradas para nuestros experimentos fueron las listas base T, TDN, junto con sus correspondientes reordenamientos resultantes de la aplicación del SRIG, y el reordenamiento de la lista base TD.

- **image08.** Para este conjunto se utilizaron listas de resultados obtenidas mediante un modelo de recuperación basado en texto (este método utiliza las descripciones textuales de las imágenes del conjunto para el indexado, y la parte textual de los tópicos para realizar la recuperación), un modelo basado en imágenes (se consideraron los resultados dados como baseline por los organizadores del CLEF, obtenidos mediante el sistema FIRE, el cual implementa un modelo de recuperación basado en contenido y realiza la recuperación mediante las imágenes ejemplo del tópico), un modelo de recuperación multimodal (en este modelo las imágenes se segmentan para extraer características visuales. Utilizando un conjunto de imágenes etiquetadas previamente, el resto de las imágenes se etiquetan automáticamente considerando un vocabulario de 222 palabras. Las etiquetas obtenidas son utilizadas para expandir la anotación manual de las imágenes. Esta expansión es utilizada para realizar el indexado de las imágenes. La parte textual de los tópicos es utilizada para realizar la recuperación.), y un método de fusión tardía de resultados de recuperación (este método utiliza una combinación lineal de los pesos dados por los diferentes modelos de recuperación, la redundancia de los elementos en las listas de resultados, y la efectividad individual de cada uno de los modelos de recuperación). De este último, dos listas de resultados fueron consideradas. Más información acerca de los métodos de recuperación y del sistema de fusión tardía pueden consultarse en [33].

Con los sistemas anteriores se recuperaron listas de 1000 elementos. Estas listas tienen el formato utilizado en la evaluación del foro TREC, el cuál se muestra a continuación:

```
251 Q0 LA030694-0247 1 0.456544 Exp
251 Q0 GH950321-000003 1 0.453528 Exp
252 Q0 GH951025-000068 2 0.694389 Exp
252 Q0 GH951227-000047 3 0.601287 Exp
253 Q0 GH950407-000095 4 0.399507 Exp
253 Q0 LA010694-0084 2 5 0.382531 Exp
```

La primera columna indica el identificador del tópico (*topicId*), la segunda indica el número de iteración (*iteration*), la tercera indica el identificador del elemento recuperado (*documentId*), la cuarta indica la posición del elemento recuperado en la lista (*rank*), la quinta indica el valor de confianza dado al elemento por el recuperador (*score*) y la sexta indica el nombre del experimento (*runId*).

Debido a que el enfoque del método desarrollado es independiente del funcionamiento interno de los sistemas de RI, no se presenta un estudio extenso de los mismos. Información detallada puede consultarse en *www.lemurproject.org*, [20] y [33].

Con lo anterior se tienen los elementos necesarios para mostrar la motivación y justificación del desarrollo de un método automático de selección previa de las listas que ayuden a mejorar la efectividad de la Fusión de Datos.

4.4. Estudio de factibilidad

La hipótesis de nuestra propuesta es que la Fusión de Datos puede ser mejorada mediante una selección previa de las listas que se fusionarán. La tabla 4.3 muestra una comparación del MAP de cada uno de las listas consideradas para cada conjunto y de la Fusión de Datos sistemática (fusión de las 5 listas), utilizando los tres métodos de fusión descritos a detalle en la sección 2.2 (MaxRSV, CombMNZ y Fuzzy Borda).

	Listas iniciales					**Fusión sistemática (5 listas)**		
Conjunto	**lista 1**	**lista 2**	**lista 3**	**lista 4**	**lista 5**	**MaxRSV**	**CombMNZ**	**Fuzzy Borda**
adhoc05	0.192	0.293	0.174	**0.300**	0.290	0.231	0.278	0.263
geoir08	0.218	0.248	**0.263**	0.210	0.218	*0.180*	0.244	**0.265**
image08	0.255	**0.292**	0.271	0.094	0.278	0.251	**0.302**	**0.317**
robust08	0.218	**0.359**	0.198	0.240	0.313	0.231	0.341	*0.165*

Tabla 4.3: Efectividad en *MAP* de las listas de resultados iniciales y la fusión sistemática

En la columna **Listas iniciales**, se presentan en negritas aquellas con los mejores resultados de recuperación, mientras que en la columna **Fusión sistemática** se muestran en negritas los métodos que lograron mejorar la efectividad de todas las listas, en cursivas se muestran aquellos con una efectividad menor al de todas las listas iniciales y los restantes son aquellos que logran superar al menos a dos listas iniciales.

Podemos observar que sólo en tres casos se logró superar *MAP* de todas las listas iniciales (geoir08 con Fuzzy Borda, e image08 con CombMNZ y Fuzzy Borda), mientas que en dos casos el *MAP* fue menor que el de todas las listas iniciales (geoir08 con MaxRSV y robust08 con Fuzzy Borda). Lo anterior muestra que, si bien la Fusión de Datos es una alternativa para mejorar los resultados de recuperación, no siempre se logra el objetivo de la misma, que como se mencionó en la sección 2.2, es enfatizar las fortalezas de las listas y disminuir sus debilidades.

Los resultados de la tabla 4.3 muestran la efectividad global de la fusión sistemática con tres métodos de fusión, sin embargo, como se mostró en la sección 1.1, no siempre un mejor resultado global determina al mejor resultado de recuperación, ya que puede suceder que las mejoras individuales de gran magnitud "disfracen" a los decrementos de efectividad individuales.

Para observar mejor el comportamiento de la Fusión de Datos sistemática, realizamos un estudio particular de la eficacia individual en cada petición, con el propósito de conocer el número de éstas en que la fusión realmente logra superar a las listas iniciales. Los resultados de este estudio se muestran en la tabla 4.4.

Podemos observar que en todos los casos, la FD sistemática no logra superar los resultados de todas las listas iniciales en más de un 50 % de las peticiones. De los casos en los que la efectividad de la fusión superó globalmente a todas las listas iniciales (tabla 4.3), sólo 6 de 24 peticiones (para el caso de geoir08 con Fuzzy Borda), 15 de 39 (para el caso de image08 con CombMNZ) y 12 de 39 (para el caso de image08 con Fuzzy Borda), logran superar la eficacia de todas las listas.

Otra situación que se presenta es el hecho de que existen peticiones en las que la fusión obtiene una efectividad inferior a la de todas las listas iniciales, lo cual contradice el objetivo de la Fusión de Datos.

Las observaciones anteriores muestran que existe oportunidad para mejorar la efectividad de la Fusión de Datos, y cómo hacerlo es el problema que atañe a esta investigación.

Nuestra hipótesis es que la Fusión de Datos puede mejorarse mediante una selec-

MaxRSV						
Conjunto	**menor**	> 1	> 2	> 3	> 4	> 5
adhoc05 (50 p.)	8	9	18	9	4	2
geoir08 (24 p.)	5	5	7	2	5	0
image08 (39 p.)	0	4	13	16	4	2
robust08 (153 p.)	28	23	36	46	16	4
CombMNZ						
Conjunto	**menor**	> 1	> 2	> 3	> 4	> 5
adhoc05 (50 p.)	0	0	19	13	9	9
geoir08 (24 p.)	0	1	8	8	1	6
image08 (39 p.)	1	1	7	12	3	15
robust08 (153 p.)	3	3	21	43	32	51
Fuzzy Borda						
Conjunto	**menor**	> 1	> 2	> 3	> 4	> 5
adhoc05 (50 p.)	0	1	24	9	7	9
geoir08 (24 p.)	1	1	5	7	4	6
image08 (39 p.)	0	1	5	9	12	12
robust08 (153 p.)	36	32	27	33	12	13

Tabla 4.4: Número de peticiones en las que la fusión supera a 1, 2, 3, 4 o a las 5 listas iniciales

ción previa de las listas a fusionar. Esto nos lleva a la pregunta ¿En qué proporción puede mejorarse la eficacia de la fusión mediante esta selección previa? Para contestar esta interrogante se diseñó el siguiente experimento:

1. Generar para cada petición la fusión de todos los posibles subconjuntos de las 5 listas iniciales.

2. Evaluar la efectividad de cada fusión generada.

3. Seleccionar como resultado para cada petición la fusión con el mejor valor de eficacia.

4. Construir una nueva lista de resultados con los resultados individuales seleccionados para cada petición.

5. Evaluar la efectividad global de la lista generada.

De esta manera, 26 fusiones por petición fueron generadas para cada conjunto (10 fusiones de 2 listas, 10 fusiones de 3 listas, 5 fusiones de 4 listas y 1 fusión de 5

listas). Después de evaluar las 26 fusiones disponibles, aquella con el mejor *MAP* fue seleccionada como resultado de recuperación. Con lo anterior, tres listas de resultados globales se obtuvieron al aplicar los métodos de fusión MaxRSV, CombMNZ y Fuzzy Borda. Los resultados obtenidos se muestran en la tabla 4.5.

MaxRSV			
Conjunto	**Fusión sistemática**	**Experimento manual**	**Ganancia**
adhoc05	0.231	0.321	**38.9 %**
geoir08	0.180	0.330	**83.3 %**
image08	0.251	0.351	**39.8 %**
robust08	0.229	0.373	**62.8 %**
CombMNZ			
Conjunto	**Fusión sistemática**	**Experimento manual**	**Ganancia**
adhoc05	0.275	0.342	**24.3 %**
geoir08	0.244	0.341	**39.7 %**
image08	0.302	0.407	**34.7 %**
robust08	0.341	0.434	**27.2 %**
Fuzzy Borda			
Conjunto	**Fusión sistemática**	**Experimento manual**	**Ganancia**
adhoc05	0.267	0.342	**28.0 %**
geoir08	0.265	0.345	**30.1 %**
image08	0.316	0.403	**27.5 %**
robust08	0.161	0.408	**153.4 %**

Tabla 4.5: Posible ganancia de una selección de listas previa a la Fusión de Datos

La tabla anterior muestra que existe una ganancia sustancial en el *MAP* sobre la fusión sistemática al realizar una correcta selección de las listas que deben fusionarse. Por otro lado, en la tabla 4.6 podemos observar para cuántas peticiones se seleccionó una fusión de 2, 3 ,4 ó 5 listas. Esta tabla muestra que, en la mayoría de los casos, fusiones de 2 listas son las que obtienen los mejores resultados. Esto se debe a que los elementos no relevantes aumentan conforme se incluyen más listas en la fusión, lo cual provoca que los métodos de fusión los incluyan en la nueva lista provocando un decremento en su efectividad.

Los análisis anteriores muestran la importancia que tendría contar con un método automático que realice una selección previa de listas a fusionar. Este método ayudaría a la Fusión de Datos a ofrecer mejores resultados de recuperación. Estos análisis también responden dos de las tres interrogantes formuladas al inicio de este capítulo,

MaxRSV				
Conjunto	**2 listas**	**3 listas**	**4 listas**	**5 listas**
adhoc05	41	7	1	1
geoir08	21	3	0	0
image08	35	4	0	0
robust08	139	13	1	0
CombMNZ				
Conjunto	**2 listas**	**3 listas**	**4 listas**	**5 listas**
adhoc05	35	12	3	0
geoir08	20	2	2	0
image08	24	7	7	1
robust08	96	39	16	2
Fuzzy Borda				
Conjunto	**2 listas**	**3 listas**	**4 listas**	**5 listas**
adhoc05	32	17	1	0
geoir08	23	1	0	0
image08	23	13	3	0
robust08	102	42	9	0

Tabla 4.6: Número de listas en las fusiones seleccionadas como mejor resultado de recuperación.

ya que muestra que un método de selección previa puede mejorar en gran medida la efectividad de la fusión sistemática y de las listas iniciales. La tercera interrogante se relaciona con los resultados obtenidos por el método de selección propuesto, el cual se presenta en detalle en el siguiente capítulo.

Capítulo 5

Fusión Dinámica de Resultados de RI

Una vez mostrada la importancia que puede tener una selección previa de las listas a fusionar, nuestra investigación continúa con el desarrollo de un método que realice este proceso de manera automática. Hemos llamado al método propuesto *Fusión Dinámica de Resultados de RI* (FDRI), el cual ofrece una selección de diferentes listas de resultados para cada petición, y además tiene las características de ser no supervisado, independiente de los juicios de relevancia y del proceso interno de los SRI; además evita la generación previa de todas las posibles fusiones de las listas iniciales. Para lograr lo anterior, nuestro método se basa en *Medidas de Calidad* de las listas de resultados, las cuales *califican* a cada lista de acuerdo a los elementos similares que comparte con las demás listas del conjunto. A continuación se muestran los detalles de las medidas de calidad propuestas para nuestro método.

5.1. Descripción de las medidas de calidad

Para asegurar las características de nuestro método, las medidas de calidad deben evitar utilizar elementos externos. Lo anterior impone una restricción muy grande ya que lo único que puede utilizarse son las peticiones, y la información contenida en las listas de resultados (identificadores de elementos, su posición en la lista y el valor de similitud que dió el recuperador al elemento). En ocasiones se cuenta con la colección de búsqueda, pero normalmente no es accesible. En este contexto, podemos definir dos clases de medidas de calidad que pueden extraerse dependiendo de los elementos con los que se cuenten. Estas clases son:

- **Medidas de calidad intrínsecas**. Califican a las listas de acuerdo a la similitud que tiene la petición con el contenido de los elementos recuperados. Requieren de las peticiones, las listas de resultados y la colección de búsqueda.

- **Medidas de calidad extrínsecas**. Califican a las listas de acuerdo a los elementos recuperados comunes entre ellas. Utilizan la redundancia y el posicionamiento de dichos elementos. Sólo requieren de las listas de resultados.

En esta investigación nos enfocamos en las medidas de calidad extrínsecas debido a que ofrecen una mayor independencia para el método de selección. Más aún, como se mostró en la sección 2.2, dos de las características utilizadas por los métodos de fusión son la redundancia y el posicionamiento (efectos *coro* y *skimming*), por lo que una de las motivaciones más fuertes de este trabajo es investigar si estas características son suficientes para hacer una selección previa de listas que ayude mejorar la efectividad en la Fusión de Datos.

En esta investigación se proponen cuatro medidas de calidad para calificar a las listas de resultados. Estas medidas son parte de las aportaciones generadas en esta investigación. A continuación se presentan sus detalles así como una justificación de su uso.

- Q_1 **- Suma de cardinalidades de intersecciones parciales del conjunto de listas.**

  Hipótesis: *cuanto más alta sea la redundancia de los elementos de una lista en las demás listas del conjunto, la calidad de la lista será mayor.*

  Esta medida calcula la calidad de una lista en función de los elementos *únicos* que contiene y en la redundancia de los elementos que comparte con el resto listas del conjunto. Los elementos *únicos* son aquellos que sólo se encuentran en la lista que se califica. Los elementos redundantes son aquellos que se encuentran en la intersección de dos listas (la lista que se califica y alguna de las otras $n-1$), elementos en la intersección de tres listas (la lista que se califica y dos de las otras $n-1$), y así sucesivamente hasta los elementos que se encuentran en la intersección de las n listas (a la cuál llamaremos *intersección completa*). La medida Q_1 considera importantes a todos los elementos de la lista que se califica, sólo que esa importancia es ponderada por el número de listas que contienen a cada elemento. Si un elemento aparece en dos o más listas de resultados, este

aporta más puntos de calidad que los elementos únicos. La diferencia con otros trabajos que también utilizan la redundancia, es que no sólo se consideran los elementos únicos y los de la intersección de las n listas disponibles, sino que se consideran intersecciones de subconjuntos de listas. Formalmente, definimos esta medida como:

$$Q_1(l, L) = \sum_{\forall d \in l} C(d, L) \tag{5.1.1}$$

donde $l = \{d_1, d_2, ..., d_n\}$ es una lista de resultados de recuperación, $L = \{l_1, l_2, .., l_m\}$ es el conjunto de listas de resultados, y $C(d, L)$ es una función que determina el número de listas en L que contienen al elemento d, la cual se define como:

$$C(d, L) = \sum_{\forall x \in L} e(d, x), \quad e(d, x) = \begin{cases} 1, & d \in x; \\ 0, & \text{en otro caso.} \end{cases} \tag{5.1.2}$$

El valor final de esta característica indica el número de elementos que se intersectan en diferentes subconjuntos de las listas, ponderados por el número de listas que los contienen. De acuerdo al efecto coro, entre más alto sea el valor de esta medida, la lista tendrá una mejor calidad.

- Q_2 **- Suma de la posición inversa de los elementos en la intersección.**

  Hipótesis: *cuanto mejor posicionados se encuentren los elementos más redundantes de una lista, la calidad de esta será mayor.*

  Con esta medida, la calidad de una lista se calcula utilizando primero el efecto coro, y después el efecto skimming. De acuerdo al efecto *coro*, un elemento que aparezca en varias listas tiene una mayor probabilidad de ser relevante, por tanto, los elementos en la intersección total del conjunto de listas son los que tienen la mayor probabilidad de ser relevantes. Además de eso, de acuerdo al efecto *skimming* los elementos que se recuperan en las primeras posiciones son los que se considera tienen una mayor similitud con la petición. Tomando en cuenta las dos observaciones anteriores, cuanto mejor posicionados se encuentren los elementos del conjunto intersección en la lista a calificar, ésta tendrá una mejor calidad. Formalmente, definimos esta medida de la siguiente manera:

$$Q_2(l, L) = \sum_{\forall d \in I} \left(\frac{1}{p(d, l)} \right) \tag{5.1.3}$$

donde

$$I = \bigcap_{\forall l \in L} l \tag{5.1.4}$$

con $l = \{d_1, d_2, ..., d_n\}$ es una lista de resultados de recuperación, $L = \{l_1, l_2, .., l_m\}$ es el conjunto de listas de resultados y $p(d, l)$ es una función que indica la posición del elemento d en la lista l. Al utilizar esta medida, si los elementos de la intersección se encuentran en las primeras posiciones de la lista evaluada, estos aportan la mayor parte del valor final de calidad, por lo que decimos que esta medida *premia* a los elementos en las primeras posiciones.

El valor final de estas características indica qué tan bien posicionados se encuentran los elementos de la intersección total en la lista evaluada. Se utiliza el inverso de las posiciones para relacionar directamente valores mayores de la medida con una mejor calidad.

- Q_3 **- Inverso de la suma de la posición de los elementos en la intersección.**

  Esta medida tiene la misma hipótesis y justificaciones que la medida Q_2, ya que utiliza los efectos *skimming* y *coro* para determinar la calidad de una lista de resultados. Sin embargo, a diferencia de Q_2, en esta medida la calidad de una lista se ve afectada si ésta contiene a los elementos del conjunto intersección en las últimas posiciones, por lo que decimos que *castiga* a los elementos en posiciones finales. Formalmente definimos esta medida de calidad como sigue:

$$Q_3(l, L) = \frac{1}{\sum_{\forall d \in I} p(d, l)} \tag{5.1.5}$$

  con $l = \{d_1, d_2, ..., d_n\}$ siendo una lista de resultados de recuperación, $L = \{l_1, l_2, .., l_m\}$ es el conjunto de listas de resultados y $p(d, l)$ es una función que indica la posición del elemento d en la lista l e I definida como en 5.1.4.

El valor final de estas características indica qué tan bien posicionados se encuentran los elementos de la intersección total en la lista evaluada. Un valor alto de esta medida indica una calidad mayor.

- Q_4 **- Suma de la posición suavizada de los elementos en la intersección total.**

  Esta medida tiene la misma hipótesis y justificaciones que la medida Q_2, sin embargo, en esta versión se utiliza un factor de suavizado en la posición de los elementos redundantes, con el fin de evitar la ventaja que otorga la posición inversa a la calidad final de las listas que contienen elementos redundantes en las primeras posiciones. Lo anterior es debido a que existen grandes saltos en los valores de la posición inversa, especialmente en las primeras posiciones. De manera formal, definimos Q_4 de la siguiente manera:

$$Q_4(l, L) = \sum_{\forall d \in I} \left(1 - \frac{ln(p(d, l))}{ln(|l|)}\right) \tag{5.1.6}$$

  donde $l = \{d_1, d_2, ..., d_n\}$ es una lista de resultados de recuperación, $L = \{l_1, l_2, .., l_m\}$ es el conjunto de listas de resultados y $p(d, l)$ es una función que indica la posición del elemento d en la lista l e I definida como en 5.1.4.

  La figura 5.1 muestra una comparación de los valores que las medidas Q_2 y Q_4 obtienen para las posiciones contínuas de una lista de 1000 elementos.

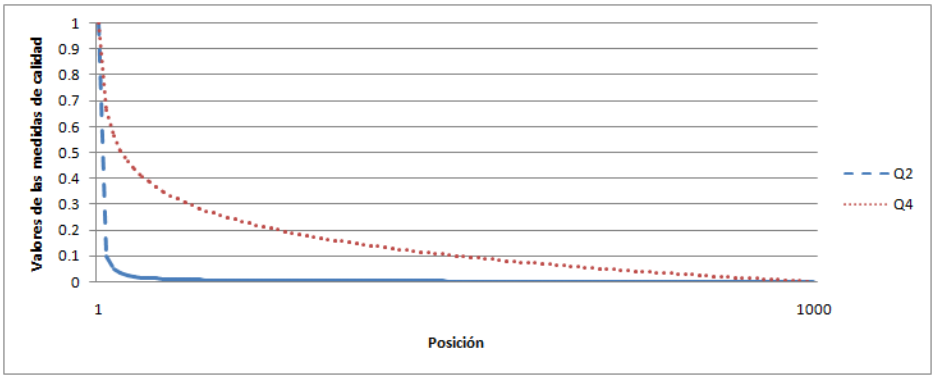

Figura 5.1: Curvas descritas por los valores obtenidos con las medidas Q_2 y Q_4.

Podemos observar que la curva que describe Q_4 desciende menos rápido que la descrita por Q_2, por lo que listas con elementos redundantes en las primeras posiciones tienen una ventaja menor al utilizar la medida de calidad Q_4.

El valor final de esta medida indica qué tan bien posicionados se encuentran los elementos de la intersección total en la lista evaluada. Un valor alto de esta medida indica una calidad mayor.

Un estudio de los valores mínimos y máximos que pueden obtener las medidas de calidad, se muestran en el apéndice C.

Ahora que se han definido las medidas de calidad para calificar a las listas de resultados de recuperación, es hora de ponerlas a prueba en el marco de trabajo de RI para conocer su eficacia.

5.2. Selección del mejor resultado de RI utilizando las medidas de calidad

En un escenario donde se cuenta con diferentes resultados de recuperación para una misma petición, obtenidos ya sea por diferentes sistemas de RI y/o diferentes estrategias de recuperación, un problema que surge de manera natural es seleccionar una sola lista de resultados para ofrecerla al usuario. Este problema es conocido como la *selección del mejor resultado de RI*.

Sean $l = \{d_1, d_2, ..., d_n\}$ una lista de resultados y $L = \{l_1, l_2, .., l_m\}$ el conjunto de listas recuperadas para una petición. Los pasos para aplicar nuestras medidas de calidad a este problema, son los siguientes:

1. Construir $\bigcup L$ (el conjunto unión de las listas de resultados) utilizando los identificadores de los d_i en las listas.

2. Construir $\bigcap L$ (el conjunto intersección de las listas de resultados) utilizando los identificadores de los d_i en las listas.

3. Calcular la calidad de las listas $Q(l)$ $\forall l \in L$, utilizando las medidas de calidad Q_1-Q_4 y los conjuntos unión o intersección, según corresponda.

4. Seleccionar como resultado final, la lista l_i tal que

$$Q(l_i) > Q(l_j)\ \forall l_j \in L, \quad i \neq j$$

Las listas seleccionadas para cada petición se utilizaron para construir una lista de resultados de recuperación global, la cual fue evaluada mediante la medida *MAP*. El no contar con información adicional sobre las listas de resultados ofreció un magnífico escenario para probar la eficacia de nuestras medidas de calidad[1].

En este experimento el número de elementos por lista fue $n = 1000$ y el número de listas consideradas fue $m = 5$. Como baseline se tomaron la lista con mejor *MAP* en cada conjunto y la fusión sistemática de las 5 listas utilizando el método CombMNZ. Los resultados para cada medida de calidad y cada conjunto se muestran en la tabla 5.1.

Conjunto	Q_1	Q_2	Q_3	Q_4	**FS CombMNZ**	**Mejor lista**	**Peor lista**
adhoc05	0.293	0.295	0.290	0.290	0.278	0.300	0.174
geoir08	0.268	0.259	0.248	0.259	0.244	0.263	0.210
image08	0.264	0.259	0.294	0.299	0.302	0.292	0.094
robust08	0.317	0.332	0.332	0.331	0.341	0.359	0.198

Tabla 5.1: Selección del mejor resultado de RI utilizando medidas de calidad

Podemos observar que las medidas de calidad propuestas logran una efectividad global muy parecida, y son capaces de construir una lista de resultados global con una eficacia equivalente a la de la mejor lista inicial. Las diferencias entre los valores de efectividad globales obtenidos por las medidas, sugieren que, dependiendo de las listas que se evalúan, alguna medida es más adecuada que las otras.

Los resultados de la tabla 5.1 muestran evidencia de que las características de redundancia y posicionamiento son útiles para seleccionar, para cada petición, un resultado equivalente con la mejor de las listas iniciales. Este hecho es de particular importancia si consideramos que en un proceso de RI cotidiano, no contamos con información acerca de la efectividad individual de diferentes resultados de recuperación.

Comparando los resultados de las medidas de calidad con la fusión sistemática de las 5 listas del conjunto, observamos que la Fusión de Datos sistemática supera, globalmente, a la selección de la mejor lista en dos de los cuatro conjuntos (image08 y robust08), además de obtener resultados equivalentes a los del baseline. Esto podría interpretarse como una ineficacia de las medidas de calidad al intentar superar a la fusión sistemática. Sin embargo, para mostrar de una manera más particular la eficacia de las medidas de calidad, se realizó un análisis en el cual se contabilizaron las

[1]Los resultados de este experimento ya han sido publicados y pueden consultarse en [4]

peticiones en las que las medidas de calidad y la fusión sistemática lograron sobrepasar el *MAP* de la lista con el mejor resultado global. Los resultados de este análisis se muestran en la figura 5.2.

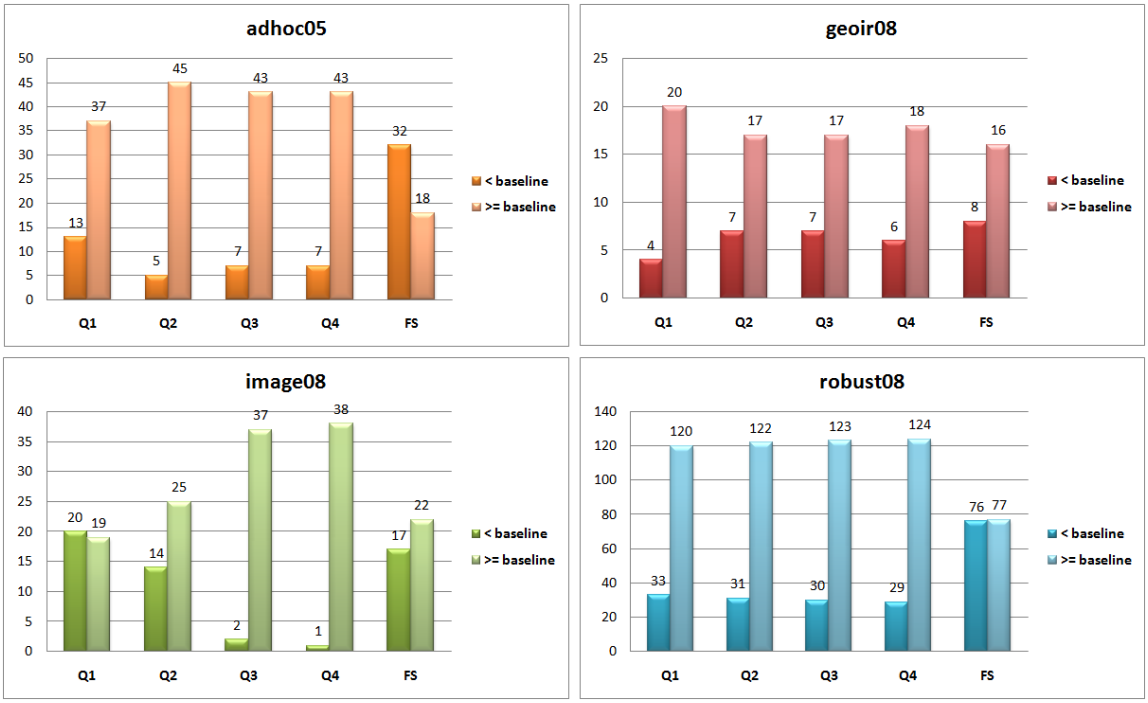

Figura 5.2: Peticiones donde la selección del mejor resultado iguala o supera al baseline

De la figura 5.2 podemos hacer las siguientes observaciones:

1. Utilizar las medidas de calidad permite seleccionar un resultado de recuperación mejor o igual al mejor resultado global en la mayoría de las peticiones.

2. Aunque la Fusión de Datos sistemática obtiene mejores resultados globales que la selección del mejor resultado, el análisis particular muestra que el número de peticiones en las que logra igualar o superar al baseline es menor.

3. Al realizar la selección de la mejor lista con las medidas de calidad se logra obtener, para la mayoría de las peticiones, una efectividad mayor a la del baseline. Sin embargo, la ganancia en *MAP* no es considerable, lo cual indica que en ocasiones se seleccionan listas con una efectividad muy baja.

4. Una de las razones por las cuales la Fusión de Datos sistemática logra un resultado global mayor es que, en cada petición, la fusión tiene la oportunidad de

sobrepasar la efectividad de todas las listas mientras que la selección del mejor resultado tiene como límite la eficacia de la mejor lista por petición.

Las observaciones anteriores indican el siguiente problema: aún cuando las medidas de calidad son eficaces para seleccionar un resultado igual o mayor al baseline, los casos en los que se selecciona algo menor influyen mucho en la evaluación global del resultado final de recuperación. Un candidato ideal para intentar resolver este problema es la Fusión de Datos. Sin embargo, como se mostró en la sección 1.1 y en los resultados del anterior análisis particular, la Fusión de Datos no siempre logra el objetivo de mejorar los resultados de recuperación. En este contexto, nuestra investigación está enfocada a mejorar los resultados obtenidos por la fusión sistemática, mediante una selección previa de las listas a fusionarse. A continuación se describe el método de Fusión Dinámica de Resultados.

5.3. Método Fusión Dinámica de Resultados de RI

El método desarrollado en este trabajo de investigación tiene como objetivo mejorar la efectividad de la Fusión de Datos mediante una selección previa de las listas más aptas para ser fusionadas, tomadas de un conjunto de resultados de recuperación disponibles. Debido la variabilidad de los resultados de recuperación disponibles, la efectividad es diferente para cada petición, por lo que las listas seleccionadas para la fusión no siempre son las mismas. Esta característica es lo que ofrece el dinamismo al método propuesto. El método requiere de los siguientes elementos:

- Un conjunto de peticiones.
- Un conjunto de resultados de recuperación obtenidos con diferentes sistemas de RI y/o estrategias de recuperación para cada petición.

Sea $l = \{d_1, d_2, ..., d_n\}$ una lista de resultados ordenados, $L = \{l_1, l_2, .., l_m\}$ el conjunto de listas recuperadas para una petición. Con lo anterior, los pasos del método FDRI son los siguientes:

1. Construir $\bigcup L$ (el conjunto unión de las listas de resultados) utilizando los identificadores de los d_i en las listas.

2. Construir $\bigcap L$ (el conjunto intersección de las listas de resultados) utilizando los identificadores de los d_i en las listas.

3. Calcular la calidad de las listas $Q(l)$ $\forall l \in L$, utilizando las medidas de calidad Q_1-Q_4 y los conjuntos unión o intersección, según corresponda.

4. Ordenar los valores $Q(l_1), \ldots, Q(l_m)$ en orden descendente.

5. Determinar S, el subconjunto de listas que serán incluidas en la fusión.

6. Realizar la fusión de las listas en S.

Debido al enfoque utilizado en el desarrollo del método, no es necesaria información adicional de los sistemas de RI, ni de la colección de búsqueda. Nuestro enfoque también tiene la característica de ser no supervisado, por lo que también evita la necesidad de juicios de relevancia y un posible re-entrenamiento para poder se aplicado a otro conjunto de datos. Más aún, el método desarrollado no necesita generar todas las posibles fusiones del conjunto inicial de listas de resultados, con lo que se logra un uso menor de recursos. Lo anterior da al método un alto grado de independencia, lo cual lo hace utilizable en cualquier contexto de RI, siempre y cuando se cuente con los elementos requeridos.

Los pasos 3 y 5 de nuestro método tienen particular importancia. El paso 3 requiere de la definición de una forma de medir la calidad de las listas de resultados. La sección 5.1 muestra los detalles de las medidas de calidad propuestas. El paso 5 necesita de una discriminación de las listas de resultados una vez medida su calidad.

En esta investigación consideramos dos estrategias para determinar el subconjunto de listas que deben incluirse en la fusión, las cuales se describen en las siguientes dos secciones.

5.4. Fusión Dinámica de Resultados de RI considerando un número fijo de listas

Este es el primer experimento elaborado para determinar la utilidad de el método FDRI para mejorar a la Fusión de Datos sistemática. En este experimento, se presentan los resultados que pueden obtener las fusiones de las mejores dos, tres o cuatro listas de resultados, según los valores de las medidas de calidad. La idea detrás de

este experimento es comprobar la tendencia observada en la tabla 4.6, la cual muestra que en el experimento manual, la fusiones de dos listas son las que tienden a obtener los mejores resultados. Teóricamente, el método FDRI debería poder capturar esta tendencia y mostrar resultados similares [2]. La descripción del experimento se muestra a continuación:

- Se considera un conjunto de $m = 5$ listas de resultados, las cuales contienen $n = 1000$ elementos recuperados, lo anterior para cada petición. Un ejemplo del contenido de las listas se muestra en la sección 4.3.
- Se aplicarán las 4 medidas de calidad a las listas de resultados, y se mostrarán los resultados obtenidos por cada una de ellas por separado.
- Los tres métodos detallados en la sección 2.2.2 (MaxRSV, CombMNZ y Fuzzy Borda), serán utilizados en el paso de fusión de las listas seleccionadas.
- En este experimento la discriminación de las listas de acuerdo a su calidad se hará de manera estática, considerando un número fijo de listas. Específicamente, se considerarán las 2, 3 y 4 mejores listas de resultados identificadas con cada medida de calidad.

Los resultados de este experimento se muestran en la sección 6.1.

5.5. Fusión Dinámica de Resultados considerando un número de listas variable

El experimento descrito anteriormente consideraba una manera simple de seleccionar las listas que deberían incluirse en la fusión, una vez medida su calidad. La estrategia era seleccionar un número fijo de listas para todas las peticiones, sin embargo como lo muestra la tabla 4.6, suele suceder que el subconjunto de listas que ofrezcan los mejores resultados en la fusión, no siempre contenga el mismo número de listas. Por esta razón se diseñó un nuevo experimento donde la selección de las listas no se realiza con número fijo, sino que depende de ciertos parámetros que tienen que ver con el valor de calidad obtenido por cada una de las listas. Esta nueva forma

[2]Los resultados de este experimento ya han sido publicados y pueden consultarse en [5]

de seleccionar el número de listas a fusionar, hace del paso 5 del método de Fusión Dinámica de Resultados (5.3), un proceso no trivial que puede ser mejorado. La idea detrás de este método de discriminación es que, al medir la calidad de las listas y ordenarlas, si existe una gran diferencia entre la lista con mejor calidad y la segunda mejor lista, entonces significaría que las listas son muy distintas entre si y por tanto no deberían fusionarse. Esta idea surge de la observación de los métodos de fusión, los cuales utilizan la redundancia (similitud entre los elementos de las listas) en su proceso. La descripción de este segundo experimento se muestra a continuación:

- Se considera un conjunto de $m = 5$ listas de resultados, las cuales contienen $n = 1000$ elementos recuperados, lo anterior para cada petición. Un ejemplo del contenido de las listas se muestra en la sección 4.3.
- Se aplicarán las 4 medidas de calidad a las listas de resultados, y se mostrarán los resultados obtenidos por cada una de ellas por separado.
- Los tres métodos detallados en la sección 2.2.2 (MaxRSV, CombMNZ y Fuzzy Borda), serán utilizados en el paso de fusión de las listas seleccionadas.
- La discriminación de las listas a fusionar se realizará mediante el monitoreo de los valores de calidad obtenidos por las listas, esto para cada petición. Los sub-pasos a seguir de este proceso (paso 5 del método FDRI), son los siguientes:

  5.1 Calcular P, el promedio de los valores $Q(l_1), \ldots, Q(l_m)$.

  5.2 Calcular las diferencias entre los valores de calidad de las listas previamente ordenadas, $d(Q(l_i), Q(l_j))$ con $1 \leq i \leq m-1$ y $j = i+1$.

  5.3 Comparar las magnitudes de las diferencias obtenidas con la magnitud del promedio, e ir incluyendo a cada listas de acuerdo a la siguiente regla:

  $$\text{Si } d(Q(l_i), Q(l_j)) \geq P \text{ entonces incluir a } l_j \text{ en la fusión.}$$

  La hipótesis que sustenta esta estrategia de discriminación es que una diferencia muy grande entre los valores de calidad de las listas (en este caso mayor al promedio de las diferencias), indicaría que dichas listas tienen un ordenamiento muy distinto de los elementos recuperados, lo cual provocaría un decremento en la efectividad de la FD si la lista con menor valor de calidad se incluyera en la fusión.

Los resultados de este experimento se muestran en la sección 6.2.

Capítulo 6

Resultados experimentales

El objetivo principal de este trabajo de investigación mejorar la efectividad de la Fusión de Datos. Con esto en mente, hemos separado los resultados obtenidos por los experimentos descritos en la secciones 5.4 y 5.5, por conjunto de datos, haciendo un análisis de los resultados obtenidos por el método FDRI con cada uno de los métodos de fusión considerados. Los conjuntos de datos utilizados en los experimentos son los descritos en la sección 4.1. A continuación se presentan los resultados experimentales obtenidos junto con un análisis de los mismos.

6.1. Resultados del método FDRI considerando un número de listas fijo

Recordando la descripción de este experimento, la forma en la que se discriminan las listas se realiza de manera estática, por lo que se ofrecen los resultados al seleccionar las primeras 2, 3 y 4 listas de resultados para realizar la fusión. En cada tabla de resultados se presentan una comparación de los resultados obtenidos por la mejor lista de resultados inicial (**ML**), los resultados de la Fusión de Datos sistemática (**FDS**), y los resultados del método FDRI al utilizar cada medida de calidad (Q_1-Q_4), considerando las 2, 3 y 4 listas con mejor calidad (**2L**, **3L** y **4L**). Se presentan los resultados de efectividad globales considerando la medida de evaluación *MAP*. Las tablas 6.1, 6.2, 6.3 y 6.4 muestran los resultados obtenidos. En negritas se presentan los resultados que lograron superar globalmente a la Fusión de Datos sistemática (fusión de las 5 listas del conjunto).

A continuación se muestra un análisis de los resultados presentados en las tablas 6.1-6.4:

adhoc05							
ML: 0.300							
		Q_1			Q_2		
Método de FD	**FDS**	**2L**	**3L**	**4L**	**2L**	**3L**	**4L**
MaxRSV	0.231	**0.248**	0.230	0.226	**0.241**	0.229	0.226
CombMNZ	0.275	**0.285**	0.271	0.270	**0.294**	**0.283**	0.269
Fuzzy Borda	0.267	**0.282**	**0.277**	**0.275**	**0.292**	**0.287**	**0.273**

		Q_3			Q_4		
Método de FD	**FDS**	**2L**	**3L**	**4L**	**2L**	**3L**	**4L**
MaxRSV	0.231	**0.244**	0.230	0.223	**0.249**	0.226	0.225
CombMNZ	0.275	**0.297**	**0.278**	0.271	**0.300**	**0.281**	0.274
Fuzzy Borda	0.267	**0.293**	**0.282**	**0.275**	**0.295**	**0.285**	**0.278**

Tabla 6.1: FDRI aplicado al conjunto adhoc05 considerando un número de listas fijo en la fusión.

geoir08							
ML: 0.263							
		Q_1			Q_2		
Método de FD	**FDS**	**2L**	**3L**	**4L**	**2L**	**3L**	**4L**
MaxRSV	0.180	**0.193**	**0.187**	0.171	**0.229**	**0.201**	0.178
CombMNZ	0.244	**0.263**	**0.262**	**0.258**	0.241	**0.267**	**0.260**
Fuzzy Borda	0.265	**0.272**	**0.271**	**0.268**	**0.270**	**0.284**	**0.286**

		Q_3			Q_4		
Método de FD	**FDS**	**2L**	**3L**	**4L**	**2L**	**3L**	**4L**
MaxRSV	0.180	**0.206**	**0.181**	0.177	**0.214**	**0.188**	0.177
CombMNZ	0.244	0.230	**0.273**	**0.261**	0.233	**0.274**	**0.261**
Fuzzy Borda	0.265	0.259	**0.287**	**0.287**	**0.266**	**0.288**	**0.286**

Tabla 6.2: FDRI aplicado al conjunto geoir08 considerando un número de listas fijo en la fusión.

- **adhoc05 (tabla 6.1)** En este conjunto de datos el método FDRI presenta una tendencia clara con todas las medidas de calidad, al utilizar el método de fusión MaxRSV: el mejor resultado se logra fusionando las dos listas con mejor calidad. Al utilizar la medida Q_1, la cual se basa sólo en la redundancia de los elementos, suele suceder que al incluir más listas en la fusión, el número de elementos no relevantes es mayor al de elementos relevantes, por lo que, en este caso, es preferible fusionar las dos listas con mejor calidad. Para las medidas Q_2, Q_3 y Q_4 sucede la misma situación con respecto a la redundancia,

image08							
ML: 0.292							
		Q_1			Q_2		
Método de FD	**FDS**	**2L**	**3L**	**4L**	**2L**	**3L**	**4L**
MaxRSV	0.251	**0.277**	**0.289**	**0.289**	**0.263**	**0.268**	**0.265**
CombMNZ	0.302	0.290	**0.303**	0.299	0.261	0.295	**0.307**
Fuzzy Borda	0.316	0.288	0.301	0.301	0.277	**0.321**	**0.332**

		Q_3			Q_4		
Método de FD	**FDS**	**2L**	**3L**	**4L**	**2L**	**3L**	**4L**
MaxRSV	0.251	**0.302**	**0.294**	**0.289**	**0.310**	**0.303**	**0.287**
CombMNZ	0.302	**0.310**	**0.310**	**0.303**	**0.333**	**0.340**	**0.330**
Fuzzy Borda	0.316	**0.319**	**0.319**	0.312	**0.341**	**0.345**	**0.335**

Tabla 6.3: FDRI aplicado al conjunto image08 considerando un número de listas fijo en la fusión.

robust08							
ML: 0.359							
		Q_1			Q_2		
Método de FD	**FDS**	**2L**	**3L**	**4L**	**2L**	**3L**	**4L**
MaxRSV	0.229	**0.276**	**0.258**	**0.238**	**0.295**	**0.259**	**0.247**
CombMNZ	0.341	0.297	0.297	0.320	0.339	0.330	0.330
Fuzzy Borda	0.161	**0.285**	**0.283**	**0.299**	**0.286**	**0.265**	**0.232**

		Q_3			Q_4		
Método de FD	**FDS**	**2L**	**3L**	**4L**	**2L**	**3L**	**4L**
MaxRSV	0.229	**0.289**	**0.265**	**0.245**	**0.288**	**0.263**	**0.246**
CombMNZ	0.341	0.334	0.324	0.323	0.335	0.325	0.324
Fuzzy Borda	0.161	**0.270**	**0.247**	**0.238**	**0.278**	**0.263**	**0.229**

Tabla 6.4: FDRI aplicado al conjunto robust08 considerando un número de listas fijo en la fusión.

pero además de eso el posicionamiento de los elementos también les afecta. Al incluir más listas de resultados, pueden existir elementos no relevantes mejor posicionados que los elementos relevantes. Globalmente, los dos mejores resultados se obtienen con las medidas Q_1 y Q_4, las cuales ofrecen una ganancia relativa sobre la fusión sistemática de un 7.35 % y 7.79 %, respectivamente. Las listas de resultados para este conjunto de datos, fusionadas con el método de fusión MaxRSV muestran una tendencia idéntica al experimento manual de la tabla 4.6, donde la efectividad decrece conforme se aumenta el número de

listas fusionadas.

Considerando en método de fusión CombMNZ, podemos observar que la eficacia de la fusión sistemática para este conjunto es superior al del método MaxRSV. El método CombMNZ se basa no sólo en la redundancia de los elementos en las listas, sino también en la posición que ocupan en las mismas. En su versión original, este método de fusión no requiere información de los valores de confianza que los sistemas de RI dan a los elementos recuperados. Lo anterior convierte al método CombMNZ en uno de los métodos de fusión más robustos y utilizados en el ámbito de Fusión de Datos. Los resultados obtenidos por el método FDRI en combinación con el método de fusión CombMNZ, muestran también una tendencia a utilizar sólo las 2 mejores listas identificadas por las medidas de calidad. En este caso, los mejores resultados se obtienen con las medidas de calidad Q_3 y Q_4, las cuales ofrecen una ganancia relativa sobre la fusión sistemática de un 8.00 % y un 9.09 %, respectivamente. En este caso observamos una correspondencia en el funcionamiento del método y las medidas de calidad, ya que los mejores resultados se obtienen considerando el posicionamiento de los elementos.

El método Fuzzy Borda es un método basado en la redundancia y en los valores de confianza que los SRIs asignan a los elementos recuperados. Este método es un buen competidor para CombMNZ, como lo muestran los resultados obtenidos por la fusión sistemática en donde Fuzzy Borda supera a CombMNZ en dos de los 4 considerados (tablas 6.1 - 6.4). Para este conjunto de datos, Fuzzy Borda con previa selección de listas via FDRI, logra superar en todos los casos a la fusión sistemática. La tendencia observada con los métodos de fusión anteriores se observa claramente, ya que los mejores resultados para las cuatro características se obtienen fusionando las dos listas con mejor calidad, y la efectividad decrece conforme se agregan más listas en la fusión. Aunque todos los resultados superan a la fusión sistemática, los dos mejores resultados globales se logran con Q_3 y Q_4, con una ganancia relativa sobre al fusión sistemática de un 9.73 % y un 10.48 %, respectivamente.

Podemos observar que el mejor método de fusión para este conjunto de listas de resultados es CombMNZ, ya que ofrece una efectividad mayor en la fusión sistemática. Al utilizar el método FDRI, se logra mejorar los desempeños globales de los tres métodos de fusión considerados, siendo el método Fuzzy Borda el más beneficiado. La medida de calidad que logra las mayores mejoras es Q_4, lo cual nos muestra evidencia de que combinar la redundancia y el posicionamiento de los elementos, para la selección de las listas a fusionar, es una buena estrategia para mejorar los resultados de la Fusión de Datos.

- **geoir08 (tabla 6.2).** En este conjunto de datos el método MaxRSV se comporta de manera similar que en el conjunto anterior, sin embargo en este caso se logra obtener un resultado global mejor al de la fusión sistemática con las 2 y 3 listas con mejor calidad. Los mejores resultados se obtienen con las medidas Q_2 y Q_4, lo cual sugiere que las listas de resultados tienen una buena redundancia y resultados de recuperación similares. Estas medidas ofrecen una ganancia relativa sobre la fusión sistemática de un 27.22 % y un 18.88 %, respectivamente.

  Con el método CombMNZ, en este conjunto observamos que tres de las cuatro medidas de calidad logran superar a la fusión sistemática utilizando 3 listas. Con 4 listas se observa un decremento en la efectividad global, mientras que al utilizar 2 listas, solamente la medida Q_1 logra una efectividad superior. El hecho de que sólo la medida Q_1 logre superar a la fusión sistemática en sus tres casos, refuerza la teoría de que la redundancia de los elementos en las listas es significativo. Sin embargo, las medidas de calidad que ofrecen los mejores resultados son Q_3 y Q_4 pero a diferencia del conjunto anterior, son tres las listas que deben fusionarse. Estas medidas ofrecen una ganancia relativa sobre la fusión sistemática de un 11.88 % y un 12.29 %, respectivamente.

  El método Fuzzy Borda muestra un comportamiento diferente al de los dos métodos de fusión anteriores ya que los mejores resultados se obtienen al fusionar 3 o 4 listas. Sólo la medida de calidad Q_1 conserva el comportamiento de preferir un número menor de listas en la fusión. Este es uno de los conjuntos donde Fuzzy Borda ofrece una mejor efectividad en la fusión sistemática que CombMNZ, por lo que el baseline para FDRI es más alto. Aún así, sólo en un

caso no se logra superar los resultados globales de la fusión sistemática. Los dos mejores resultados logran una ganancia relativa de un 8.30 % y un 8.67 %, correspondientes a las medidas de calidad Q_3 y Q_4, respectivamente.

En este conjunto la fusión sistemática con el método MaxRSV tiene una efectividad muy baja, lo cual indica que aunque la redundancia es buena en el conjunto de listas, los elementos relevantes no se encuentran en las primeras posiciones. El uso del método FDRI logra aumentar la eficacia del método en una muy buena proporción, demostrando que la selección de listas es efectiva. Con el método de fusión CombMNZ, los mejores resultados con FDRI se obtienen al fusionar un número mayor de listas; esto indica que el posicionamiento de los elementos relevantes se beneficia con la redundancia en las listas, lo cual es la base de este método de fusión. El método Fuzzy Borda es el que obtiene la mayor efectividad mediante la fusión sistemática. Ya que este método se basa en el valor de confianza de los elementos en las listas, la efectividad de este método se relaciona con una distribución de valores de confianza parecidos en las diferentes listas de resultados. Aún con un baseline más alto, FDRI logra superar el a la fusión sistemática. Una vez más observamos que las medidas de calidad son efectivas para mejorar la Fusión de Datos sistemátia, de manera particular Q_4 muestra ser útil en todos los casos analizados hasta ahora.

- **image08 (tabla 6.3).** En este conjunto de datos es donde se obtienen los mejores resultados tanto con la fusión sistemática, como con el método FDRI. En este conjunto los baselines que impone la fusión sistemática son muy altos: con MaxRSV se obtiene el resultado más cercano a la mejor lista inicial, mientras que con CombMNZ y Fuzzy Borda, la fusión sistemática logra superar la efectividad de la mejor lista inicial. Aún con esto, el método FDRI muestra una vez más su utilidad.

  Considerando el método MaxRSV, el método FDRI logra superar el baseline con todas las medidas de calidad, en todos los casos. Esto es un hecho importante, y muestra indicios de que tanto la redundancia de los elementos relevantes como su posición en las listas, son buenas. En particular, los mejores resultados se obtienen con las medidas Q_3 y Q_4, las cuales obtienen una

ganancia relativa de un 20.31 % y un 23.50 %, respectivamente. Estas dos medidas de calidad conservan la tendencia de obtener los mejores resultados con las dos listas mejor calificadas, y conforme se incluyen más listas en la fusión, la efectividad decrece. Más aún, estas dos medidas de calidad con el método MaxRSV logran superar los resultados de recuperación de la mejor lista inicial.

La fusión sistemática con el método CombMNZ logra superar a la mejor lista inicial. Esto impone un baseline muy alto para el método FDRI, lo cual se ve reflejado en los resultados obtenidos. Las medidas Q_1 y Q_2 logran superar al baseline sólo considerando 3 listas, aunque la ganancia es mucho menor que en los conjuntos anteriores. Las medidas Q_3 y Q_4 logran superar al baseline en todos los casos, obteniendo los mejores resultados para este método de fusión al considerar 3 listas de resultados. La mejor ganancia relativa obtenida por estas dos características es de un 10.26 % y un 12.58 %, respectivamente.

Con el método Fuzzy Borda la situación es aún más retadora, ya que este es el otro conjunto donde este método de fusión obtiene mejores resultados que CombMNZ en la fusión sistemática. Por esta razón, el baseline para FDRI es aún mayor. Sin embargo, una vez más se muestra la utilidad de nuestro método de selección ya que es capaz de mejorar los resultados del baseline, aún cuando éste es 8.21 % relativamente más alto que la mejor lista inicial. Los mejores resultados se obtienen con las medidas Q_2 y Q_4, las cuales logran una ganancia relativa sobre la fusión sistemática de un 5.06 % y un 9.17 %, respectivamente.

En este conjunto de datos, el método de selección FDRI muestra su utilidad al mejorar los resultados de las fusiones sistemáticas, aún cuando éstas obtienen desempeños muy altos, incluso mayores a la mejor lista inicial. Debido a las características en las que se basa nuestro método, el hecho de buena efectividad en este conjunto indica que las listas iniciales tienen una redundancia de elementos relevantes muy buena, además de que el posicionamiento de dichos elementos también es bueno, al menos en la mayoría de las listas. Para este conjunto, el método Fuzzy Borda muestra ser el más adecuado. Con respecto a las medidas

de calidad, Q_4 sigue mostrando ser la mejor de todas.

- **robust08 (tabla 6.4).** Este es el conjunto en el que la fusión sistemática obtiene los desempeños más bajos, con todos los métodos de fusión.

  Con el método MaxRSV, FDRI logra superar el resultado del baseline en todos los casos. La tendencia es la observada normalmente: los mejores resultados se logran fusionando las dos listas con mejor calidad, y va decrementándose conforme se incluyen más listas. En la descripción de los conjuntos de tópicos (sección 4.1), se menciona que este conjunto es particularmente difícil, ya que muchos tópicos contiene un número muy bajo de elementos relevantes, lo cual hace que incluir muchas listas en la fusión incremente en gran proporción el número de elementos no relevantes. Los mejores resultados ofrecen una ganancia relativa del 28.82 % con la medida Q_2 y de un 26.20 % al utilizar Q_3.

  Al utilizar el método CombMNZ en este conjunto, la fusión sistemática logra una efectividad global muy cercano al de la mejor lista inicial. El método FDRI obtiene resultados cercanos a la fusión sistemática, pero no logra superarlos. El hecho de que las peticiones tengan pocos elementos relevantes en este conjunto de datos, hace que los elementos redundantes sean, en su mayoría, no relevantes. Dado que las medidas de calidad se basan en el posicionamiento de los elementos comunes entre las listas de resultados, el hecho de que estos elementos sean en su mayoría no relevantes afecta seriamente la calificación de calidad obtenido por las medidas. Esto causa que los mejores resultados se obtengan con pocas listas en la fusión, ya que existen menos elementos "basura". Al utilizar FDRI, las medidas de calidad que ofrecen los mejores resultados son Q_2 y Q_4.

  Fuzzy Borda muestra un comportamiento radicalmente opuesto al observado hasta ahora. La efectividad de la fusión sistemática con este método baja drásticamente. Esto se debe a la diferencia en los valores de confianza de las listas de resultados y a los malos resultados de recueperación de alguna de las listas de iniciales. Este caso muestra claramente la problemática que existe cuando se incluye en la fusión una lista con resultados inusualmente bajos (efecto caballo negro, sección 2.2.2). En este caso, la utilidad del método FDRI es clara, ya

que seleccionar las listas más aptas puede mejorar a la fusión sistemática con Fuzzy Borda, hasta en un 85.71 %, correspondiente a la medida de calidad Q_1. El segundo mejor resultado se obtiene con Q_2 (77.63 % de ganancia relativa).

Los resultados obtenidos en este conjunto de datos, muestran que las listas de resultados contienen elementos redundantes que en su mayoría no son relevantes. Esto provoca que los resultados de FDRI con el método CombMNZ no logre superar al baseline, ya que los elementos que se utilizan son los de la intersección total de las listas iniciales, de los cuales, sólo una proporción muy pequeña son relevantes. Con respecto a las medidas de calidad, aunque Q_4 no obtiene los mejores resultados, su efectividad es muy parecido al de las demás medidas.

El análisis anterior sugiere ciertas características acerca de la redundancia de elementos en las listas de resultados. Para reforzar estas observaciones, la tabla 6.5 muestra un análisis de las cardinalidades de los conjuntos unión e intersección de las listas iniciales, y de los elementos relevantes que contienen dichos conjuntos. Los datos mostrados corresponden al promedio de los valores obtenidos por petición.

En esta tabla podemos observar que el conjunto **robust08** tiene la cantidad más baja de elementos relevantes en la intersección, comparada con la cardinalidad de la intersección. La desviación estándar muestra además que las cantidades de elementos relevantes en la intersección para cada petición son muy variadas. Además de eso, dentro del análisis realizado se identificaron para este conjunto, 7 casos en los que la intersección fue vacía y 14 casos donde la intersección no contenía ningún elemento relevante (para los otros conjuntos no existieron intersecciones vacías y se identificó sólo un caso donde la intersección no contenía elementos relevantes). Por otro lado, el conjunto **image08** contiene más elementos relevantes en la intersección, lo cual permite que nuestras medidas de calidad logren buenos desempeños al aplicar el método FDRI a este conjunto.

Además de los resultados globales, se realizó un análisis de la efectividad obtenida en cada petición, para determinar si el uso del método FDRI logra superar de manera individual a los resultados de la fusión sistemática. En este análisis se contabilizaron las peticiones en las que la fusión sistemática y el método FDRI con las diferentes medidas de calidad, logran superar a la mejor lista inicial. Los resultados se muestran en las figuras 6.1 a 6.12.

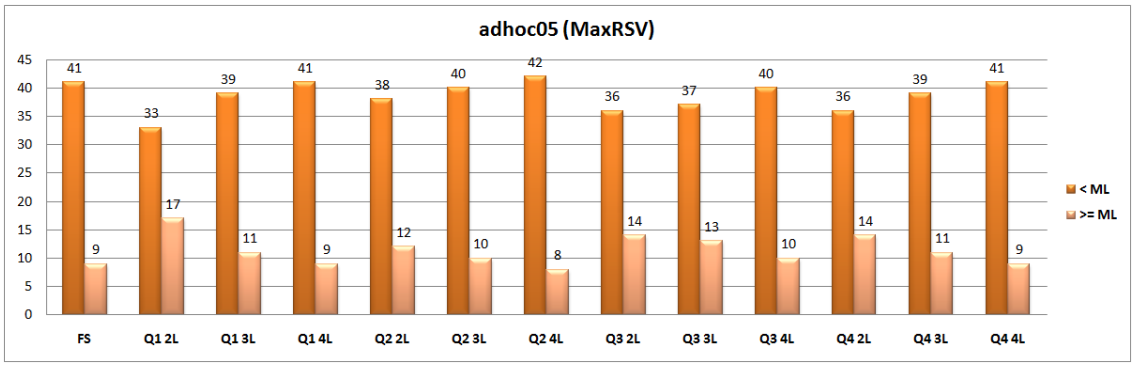

Figura 6.1: Análisis de la efectividad individual del método FDRI con listas fijas y la fusión sistemática (FS) en el conjunto adhoc05, utilizando el método de fusión MaxRSV.

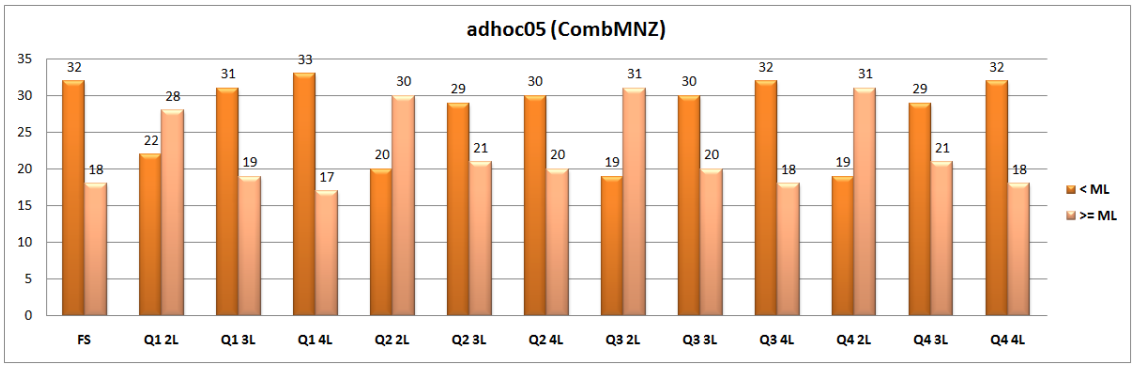

Figura 6.2: Análisis de la efectividad individual del método FDRI con listas fijas y la fusión sistemática (FS) en el conjunto adhoc05 utilizando, el método de fusión CombMNZ.

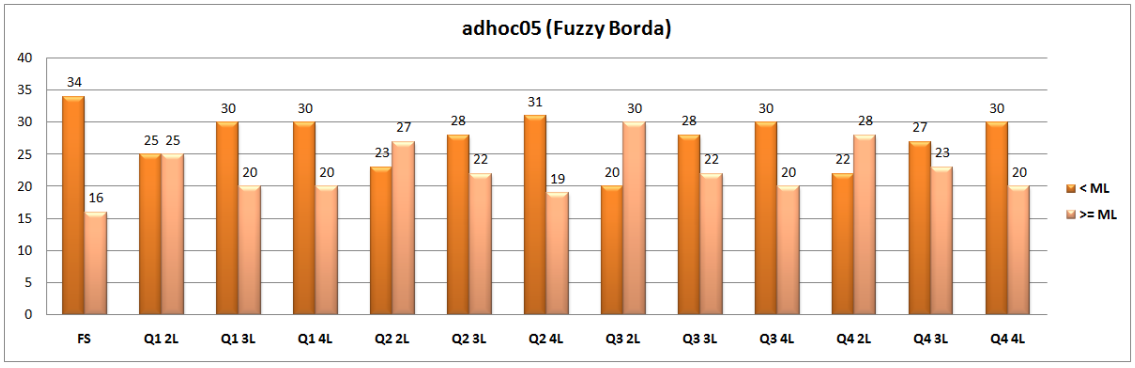

Figura 6.3: Análisis de la efectividad individual del método FDRI con listas fijas y la fusión sistemática (FS) en el conjunto adhoc05, utilizando el método de fusión Fuzzy Borda.

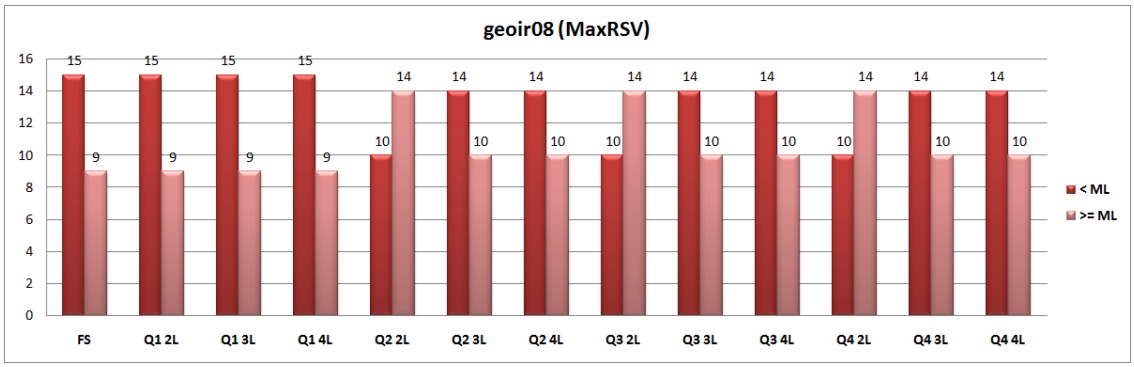

Figura 6.4: Análisis de la efectividad individual del método FDRI con listas fijas y la fusión sistemática (FS) en el conjunto geoir08, utilizando el método de fusión MaxRSV.

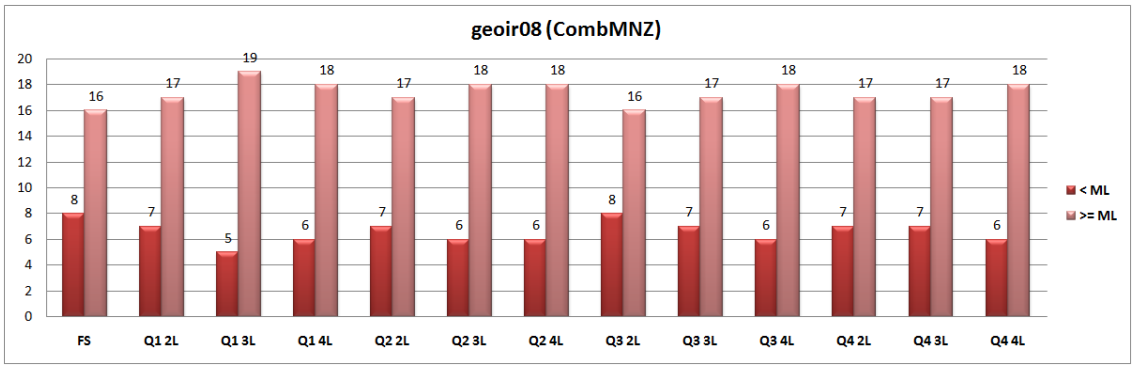

Figura 6.5: Análisis de la efectividad individual del método FDRI con listas fijas y la fusión sistemática (FS) en el conjunto geoir08, utilizando el método de fusión CombMNZ.

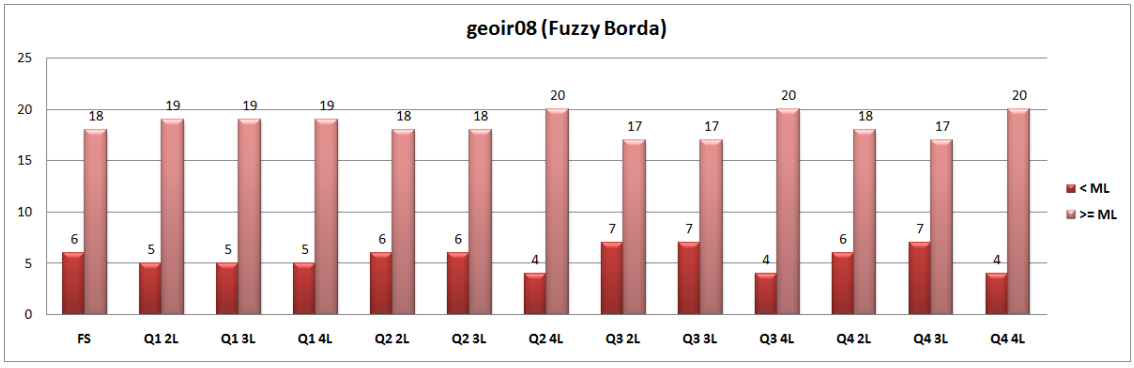

Figura 6.6: Análisis de la efectividad individual del método FDRI con listas fijas y la fusión sistemática (FS) en el conjunto geoir08, utilizando el método de fusión Fuzzy Borda.

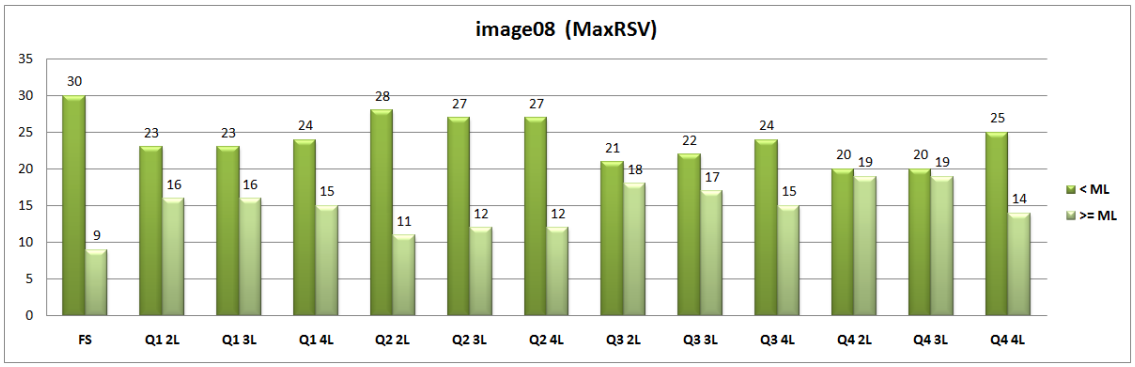

Figura 6.7: Análisis de la efectividad individual del método FDRI con listas fijas y la fusión sistemática (FS) en el conjunto image08, utilizando el método de fusión MaxRSV.

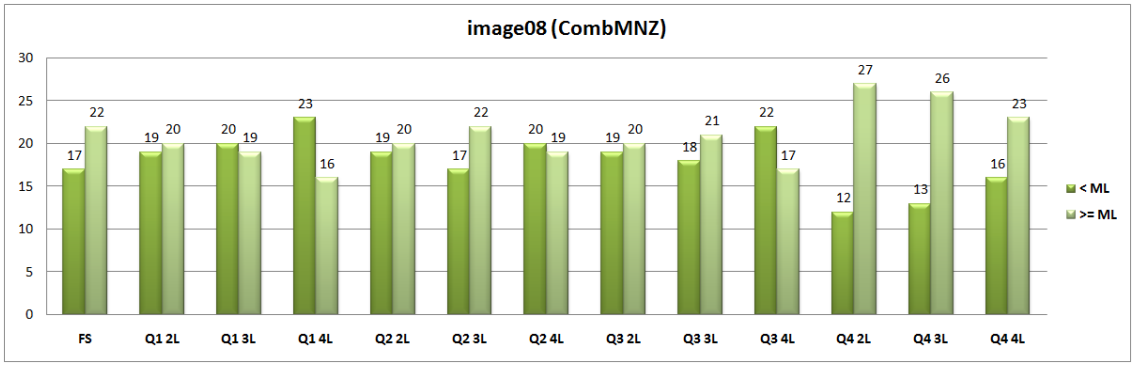

Figura 6.8: Análisis de la efectividad individual del método FDRI con listas fijas y la fusión sistemática (FS) en el conjunto image08, utilizando el método de fusión CombMNZ.

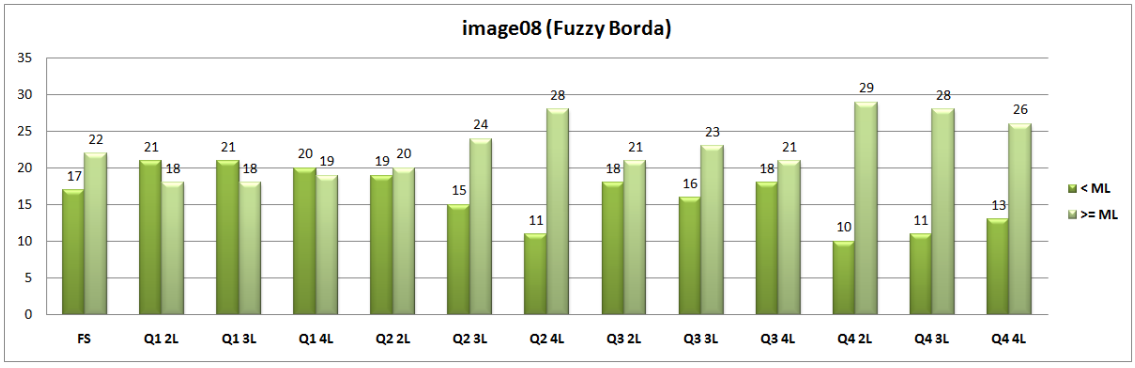

Figura 6.9: Análisis de la efectividad individual del método FDRI con listas fijas y la fusión sistemática (FS) en el conjunto image08, utilizando el método de fusión Fuzzy Borda.

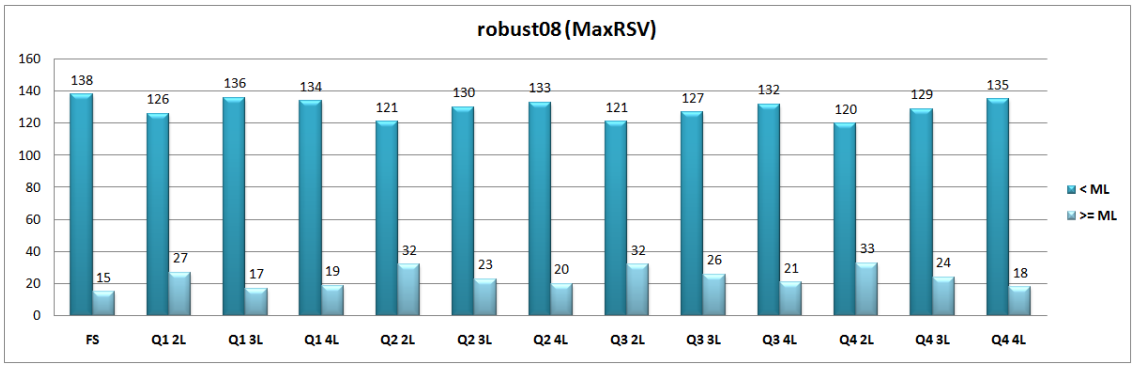

Figura 6.10: Análisis de la efectividad individual del método FDRI con listas fijas y la fusión sistemática (FS) en el conjunto robust08, utilizando el método de fusión MaxRSV.

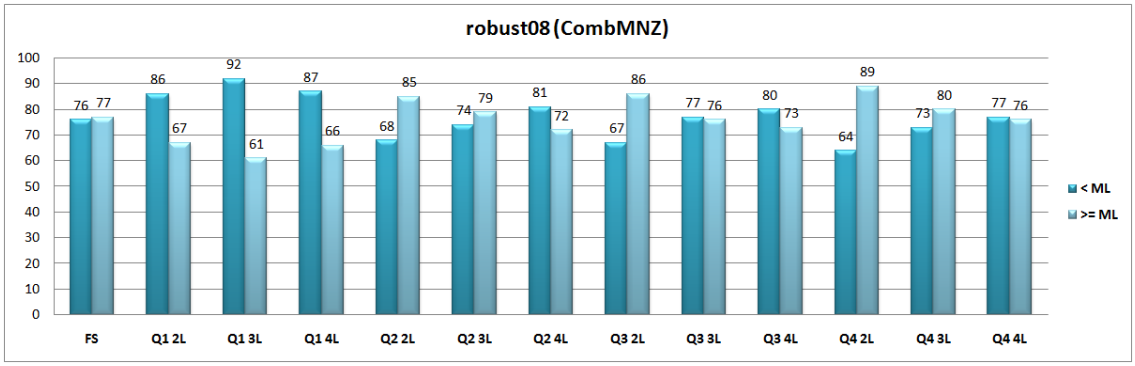

Figura 6.11: Análisis de la efectividad individual del método FDRI con listas fijas y la fusión sistemática (FS) en el conjunto robust08, utilizando el método de fusión CombMNZ.

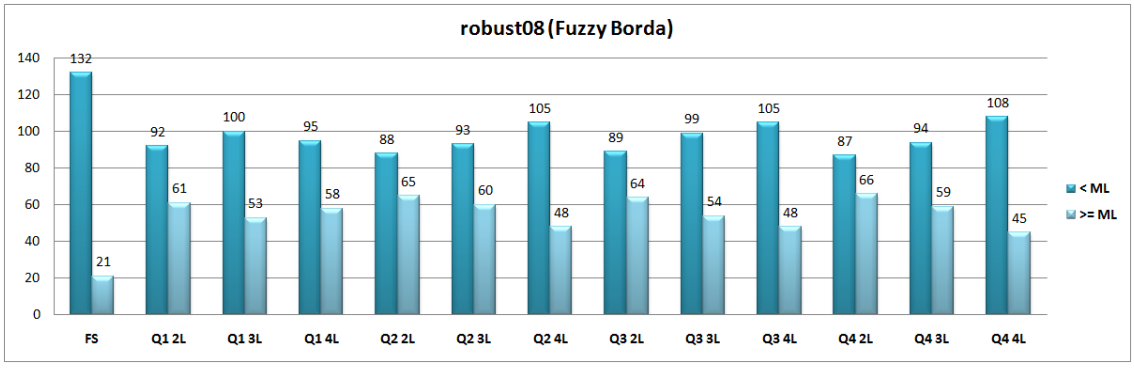

Figura 6.12: Análisis de la efectividad individual del método FDRI con listas fijas y la fusión sistemática (FS) en el conjunto robust08, utilizando el método de fusión Fuzzy Borda.

adhoc05				
Conjunto	**Mínimo**	**Máximo**	**Promedio**	**Desv. est.**
Unión	1091	2773	1918.58	290.19
Relevantes en la unión	3	200	37.08	36.74
Intersección	18	905	347.10	159.27
Relevantes en la intersección	0	115	27.76	25.52
geoir08				
Conjunto	**Mínimo**	**Máximo**	**Promedio**	**Desv. est.**
Unión	1080	1924	1473.42	201.67
Relevantes en la unión	0	105	28.37	28.34
Intersección	286	931	565.58	165.09
Relevantes en la intersección	0	82	22.75	23.33
image08				
Conjunto	**Mínimo**	**Máximo**	**Promedio**	**Desv. est.**
Unión	2046	3001	2396.54	202.95
Relevantes en la unión	18	173	55.38	30.63
Intersección	2	176	75.07	36.78
Relevantes en la intersección	0	49	15.08	10.81
robust08				
Conjunto	**Mínimo**	**Máximo**	**Promedio**	**Desv. est.**
Unión	1633	3382	2703.63	374.48
Relevantes en la unión	1	196	26.41	30.34
Intersección	0	587	137.64	121.21
Relevantes en la intersección	0	115	15.20	19.30

Tabla 6.5: Detalles de los elementos relevantes en la unión y la intersección de las listas iniciales.

Las figuras anteriores muestran que el método FDRI logra superar en la mayoría de las casos, el número de peticiones en las que se obtienen mejores resultados que aquellos de la mejor lista inicial, para cada conjunto. Debemos notar que la medida Q_4 es consistente en sus resultados para todos los conjuntos, ya que puede observarse que obtiene los mejores resultados en la mayoría de los casos.

Con este experimento pudimos comprobar que los resultados del método FDRI van acorde a la tendencia del experimento manual de la tabla 4.6, donde se muestra que en la mayoría de los casos incluir sólo dos listas ofrece la major efectividad.

6.2. Resultados del método FDRI considerando un número de listas variable

El análisis manual de la tabla 4.6 muestra que, en ocasiones, el subconjunto de listas que ofrecen los mejores resultados al fusionarse, no siempre contiene el mismo número de listas. Este experimento considera una selección de listas variable, en la cual pueden incluirse 2, 3, 4 listas en la fusión, e incluso puede suceder que sólo se seleccione la lista con el mejor valor de calidad como resultado final para una petición. Las tablas 6.6, 6.7 y 6.8 muestran los resultados obtenidos. En cada tabla se presenta una comparación de los resultados obtenidos por la mejor lista de resultados inicial (**ML**), los resultados de la fusión sistemática (**FS**), y los resultados del método de Fusión Dinámica de Resultados al utilizar cada medida de calidad (Q_1-Q_4) para la selección variable de listas. Los resultados globales se presentan con la medida de evaluación *MAP*. En negritas se presentan los resultados que lograron superar globalmente a la Fusión de Datos sistemática (fusión de las 5 listas del conjunto).

MaxRSV						
Conjunto	**ML**	**FS**	Q_1	Q_2	Q_3	Q_4
adhoc05	0.300	0.231	**0.279**	**0.280**	**0.259**	**0.274**
geoir08	0.263	0.180	**0.270**	**0.238**	**0.195**	**0.217**
image08	0.292	0.251	**0.264**	**0.261**	**0.295**	**0.306**
robust08	0.359	0.229	**0.303**	**0.320**	**0.288**	**0.308**

Tabla 6.6: FDRI con el método MaxRSV considerando un número de listas variable.

CombMNZ						
Conjunto	**ML**	**FS**	Q_1	Q_2	Q_3	Q_4
adhoc05	0.300	0.275	0.273	**0.300**	**0.296**	**0.299**
geoir08	0.263	0.244	**0.269**	**0.259**	0.232	**0.258**
image08	0.292	0.302	0.264	0.274	0.295	**0.323**
robust08	0.359	0.341	0.313	0.327	0.320	0.334

Tabla 6.7: FDRI con el método CombMNZ considerando un número de listas variable.

En las tablas anteriores observamos lo siguiente:

Con el método de fusión MaxRSV, FDRI considerando un número de listas variable permite obtener resultados estables y muy parecidos a los mostrados en las

Fuzzy Borda						
Conjunto	**ML**	**FS**	Q_1	Q_2	Q_3	Q_4
adhoc05	0.300	0.267	**0.270**	**0.297**	**0.293**	**0.294**
geoir08	0.263	0.265	**0.269**	0.262	0.264	0.263
image08	0.292	0.316	0.264	0.282	0.304	**0.329**
robust08	0.359	0.161	**0.303**	**0.301**	**0.272**	**0.292**

Tabla 6.8: FDRI con el método Fuzzy Borda considerando un número de listas variable.

tablas 6.1 - 6.4. En todos los conjuntos, las medidas de calidad logran superar a la fusión sistemática. Para los conjuntos **adhoc05**, **geoir08** y **robust08**, FDRI logra mejores ganancias relativas sobre la fusión sistemática (hasta un 21.21 % (Q_2), 50.00 % (Q_1) y 39.73 % (Q_2), respectivamente). Por otro lado, para el conjunto **image08** esta estrategia parece no ser adecuada, ya que, aunque se logra superar a la fusión sistemática y a la mejor lista individual, la ganancia es menor que la obtenida con la selección de un número de listas fijo.

Con el método CombMNZ, la estrategia implementada no muestra ningún tipo de ganancia, ya que aunque se logra superar en varios casos a la fusión sistemática y a la mejor lista inicial, la efectividad obtenida por FDRI considerando un número de listas variable, es menor que al fijar el número de listas en la fusión.

Los resultados de Fuzzy Borda con FDRI considerando un número de listas variable, muestran que esta estrategia beneficia a la fusión en el conjunto **robust08**, ya que se logra una ganancia relativa de hasta un 88.19 % con la medida Q_1. Este resultado es superior a los obtenidos con la selección de un número fijo de listas. Sin embargo, para los otros tres conjuntos la estrategia no muestra ganancias significativas.

Las figuras 6.13 - 6.24 muestran el número de peticiones que superan a la mejor lista inicial.

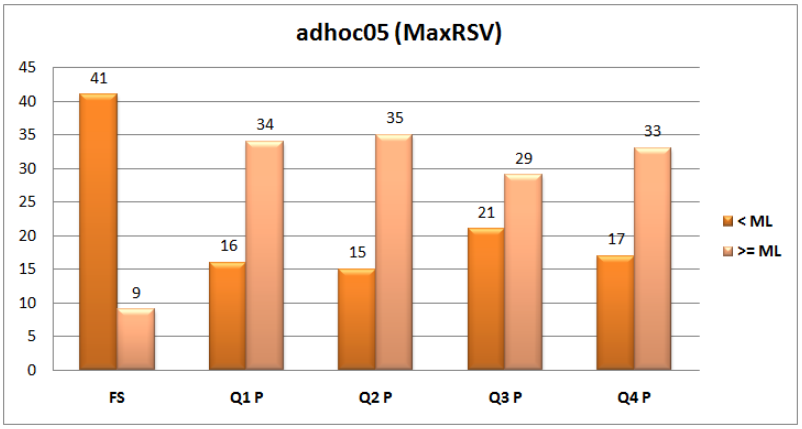

Figura 6.13: Análisis de la efectividad individual del método FDRI con listas variables y la fusión sistemática (FS) en el conjunto adhoc05, utilizando el método de fusión MaxRSV.

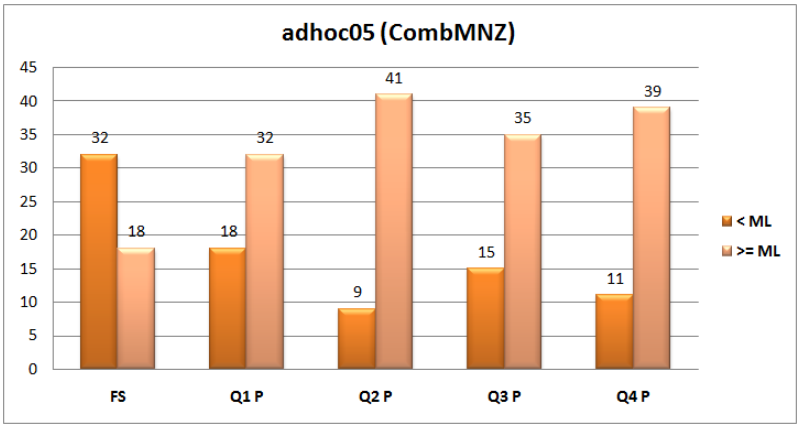

Figura 6.14: Análisis de la efectividad individual del método FDRI con listas variables y la fusión sistemática (FS) en el conjunto adhoc05, utilizando el método de fusión CombMNZ.

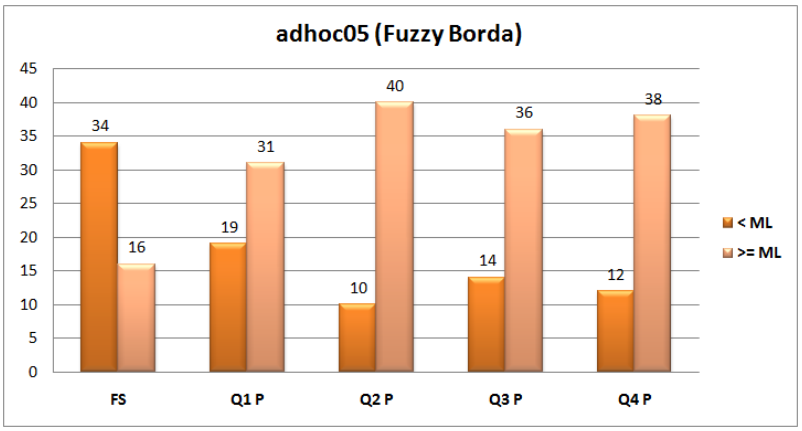

Figura 6.15: Análisis de la efectividad individual del método FDRI con listas variables y la fusión sistemática (FS) en el conjunto adhoc05, utilizando el método de fusión Fuzzy Borda.

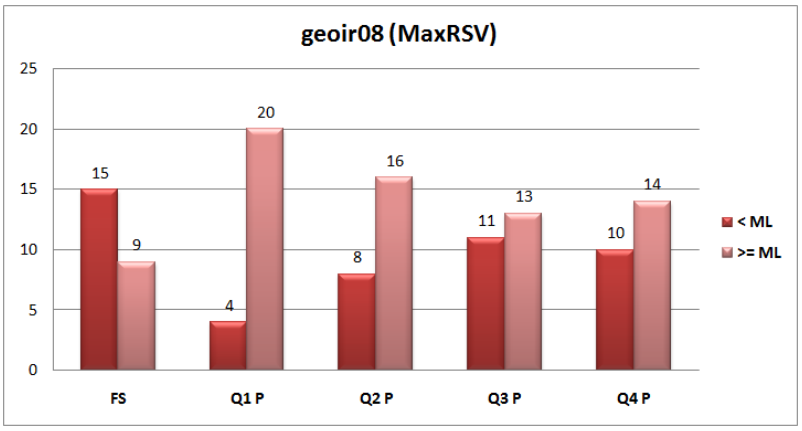

Figura 6.16: Análisis de la efectividad individual del método FDRI con listas variables y la fusión sistemática (FS) en el conjunto geoir08, utilizando el método de fusión MaxRSV.

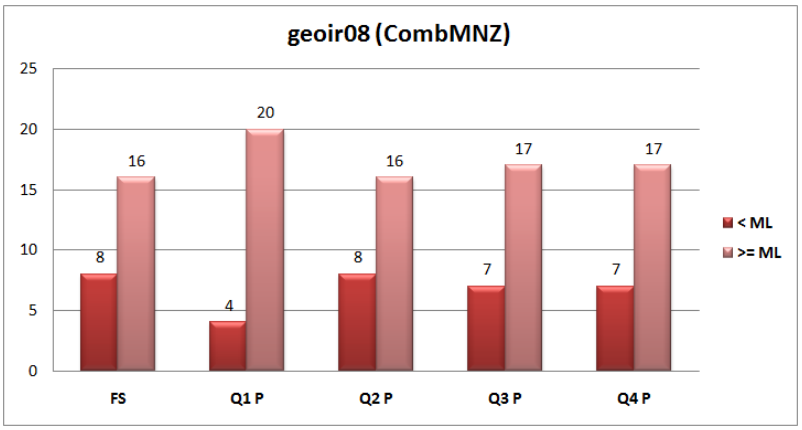

Figura 6.17: Análisis de la efectividad individual del método FDRI con listas variables y la fusión sistemática (FS) en el conjunto geoir08, utilizando el método de fusión CombMNZ.

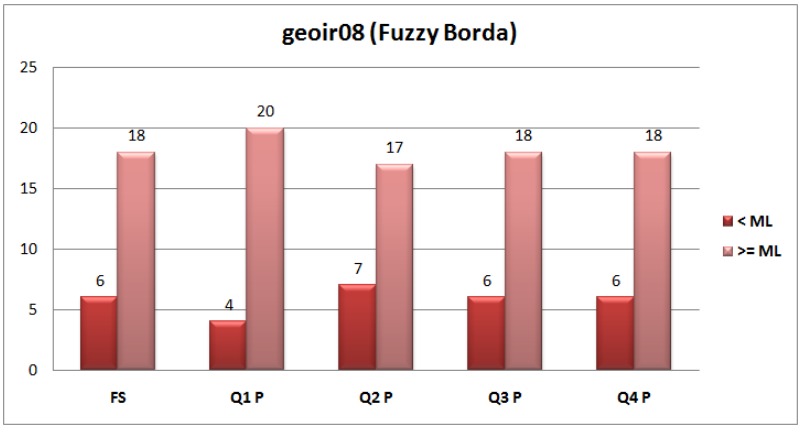

Figura 6.18: Análisis de la efectividad individual del método FDRI con listas variables y la fusión sistemática (FS) en el conjunto geoir08, utilizando el método de fusión Fuzzy Borda.

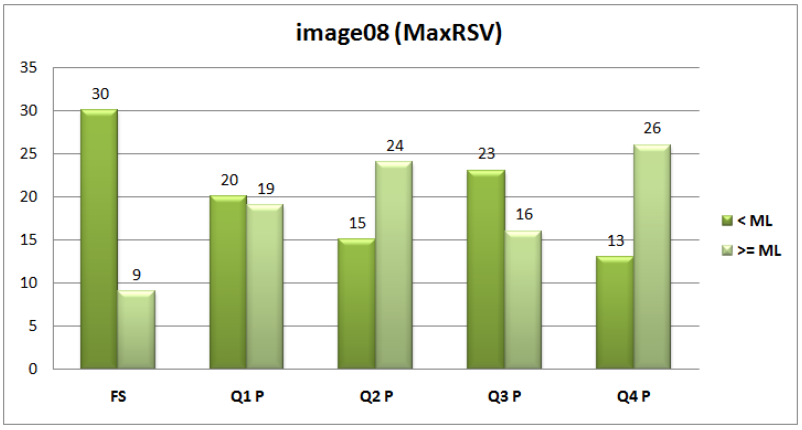

Figura 6.19: Análisis de la efectividad individual del método FDRI con listas variables y la fusión sistemática (FS) en el conjunto image08, utilizando el método de fusión MaxRSV.

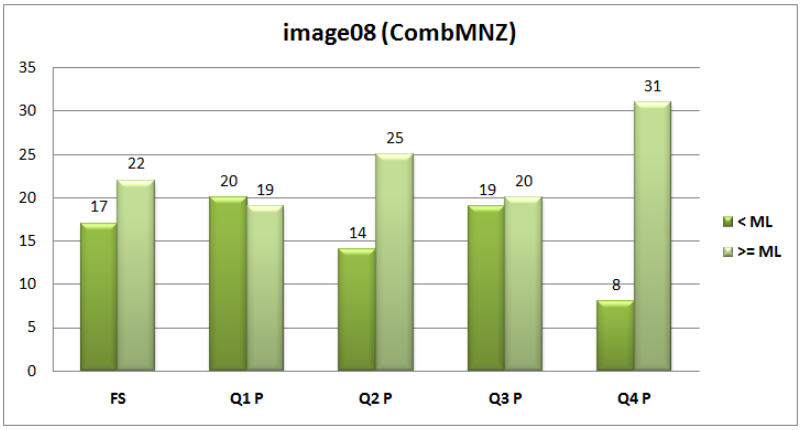

Figura 6.20: Análisis de la efectividad individual del método FDRI con listas variables y la fusión sistemática (FS) en el conjunto image08, utilizando el método de fusión CombMNZ.

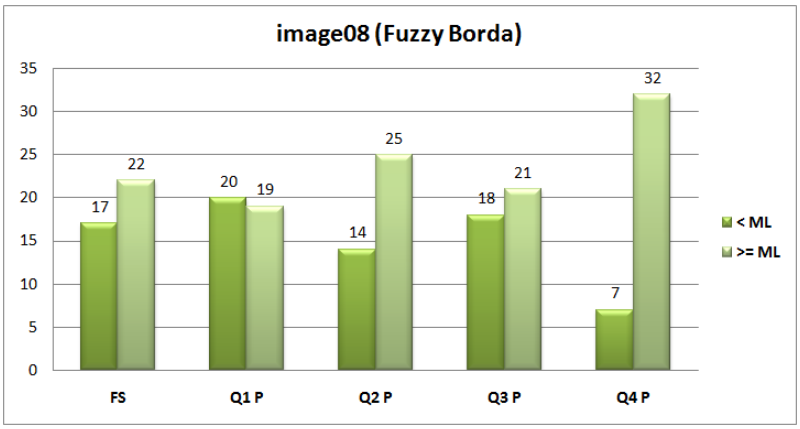

Figura 6.21: Análisis de la efectividad individual del método FDRI con listas variables y la fusión sistemática (FS) en el conjunto image08, utilizando el método de fusión Fuzzy Borda.

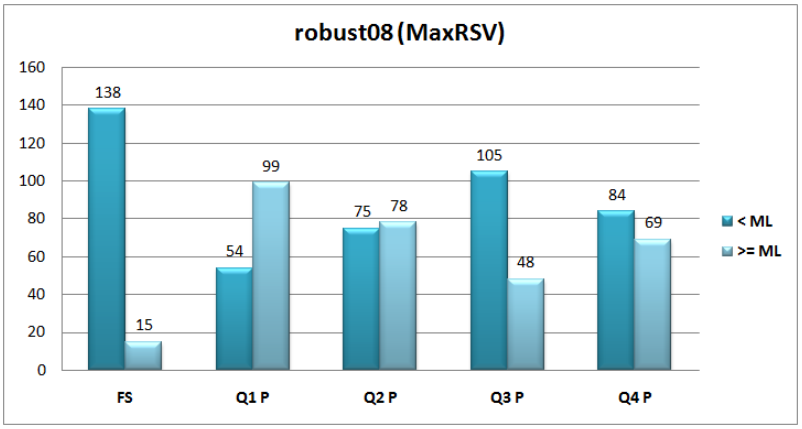

Figura 6.22: Análisis de la efectividad individual del método FDRI con listas variables y la fusión sistemática (FS) en el conjunto robust08, utilizando el método de fusión MaxRSV.

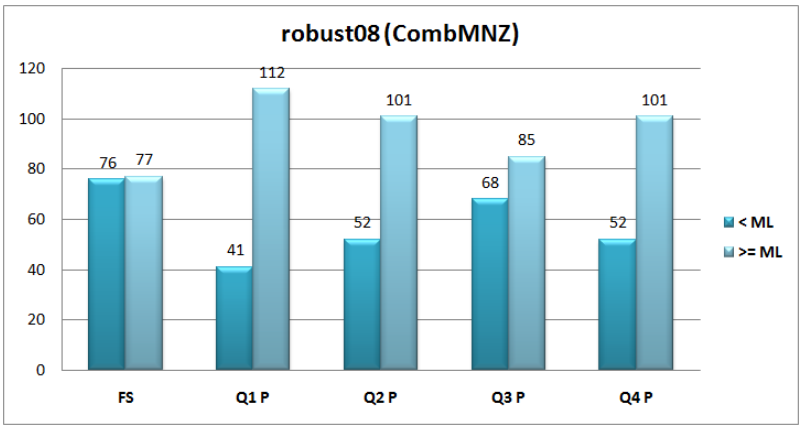

Figura 6.23: Análisis de la efectividad individual del método FDRI con listas variables y la fusión sistemática (FS) en el conjunto robust08, utilizando el método de fusión CombMNZ.

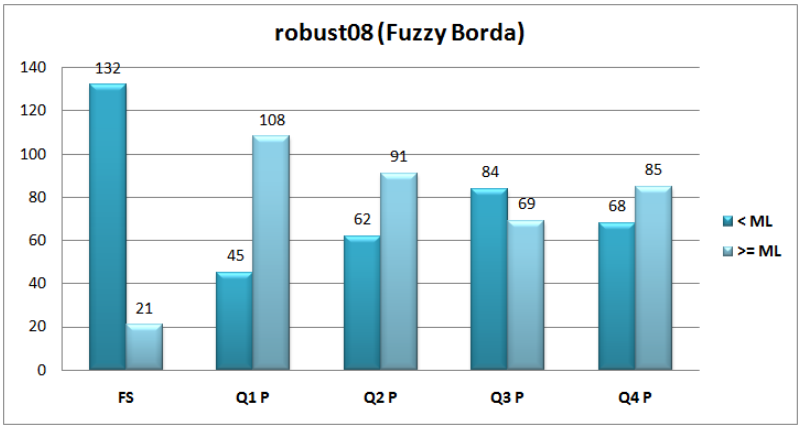

Figura 6.24: Análisis de la efectividad individual del método FDRI con listas variables y la fusión sistemática (FS) en el conjunto robust08, utilizando el método de fusión Fuzzy Borda.

Los análisis individuales muestran que, para el conjunto **adhoc05**, la estrategia utilizada hace que los tres métodos de fusión logren que un número mayor de peticiones supere el la efectividad de la mejor lista inicial. Para este conjunto de datos, la estrategia implementada en este experimento fue útil. Aunque los resultados globales fueron parecidos a los del primer experimento, los resultados individuales se ven altamente beneficiados. En el conjunto **geoir08**, los resultados del método MaxRSV mejoran individualmente, y los métodos CombMNZ y Fuzzy Borda obtienen resultados individuales similares a los del experimento anterior. Para el conjunto **image08**, se observa una mejora de los resultados individuales con los tres métodos de fusión, especialmente con la medida Q_4. En el conjunto **robust08** observamos una mayor cantidad de peticiones que superan a la mejor lista inicial en casi todos los casos.

En general, la estrategia de selección de un número variable de listas a fusionar, basada en el promedio de las diferencias de los valores de calidad de las listas, logra mejorar los resultados de manera individual, aunque de manera global ofrece una eficacia menor. Este caso remite al dilema de decidir qué es mejor, pocos resultados con una alta efectividad, o muchos resultados con una eficacia aceptable.

6.3. Validación estadística de los resultados

Los resultados experimentales muestran que el método FDRI es capaz de mejorar la efectividad de la fusión sistemática tanto de manera global (MAP) como en el número de peticiones que mejoran individualmente. Sin embargo, la tercera interrogante planteada al inicio del capítulo 4, acerca de la significancia de los resultados, continúa pendiente. Para contestar esta interrogante aplicamos un método de validación estadística al mejor resultado global, con el propósito de determinar si el aumento en la efectividad logrado por nuestro método es realmente significativo.

El análisis de significancia estadística realizado a los resultados de FDRI, se realizó tomando como base la efectividad de la fusión sistemática, debido a que el principal objetivo del método desarrollado es mejorar los resultados de este tipo de fusión. Adicionalmente, se incluye un análisis de significancia estadística utilizando como base el promedio de los desempeños de las listas iniciales, ya que en una situación donde no se cuente con los juicios de relevancia ni información acerca de las listas de resultados iniciales, tendría que seleccionarse un resultado final con un proceso un tanto aleatorio. Este análisis adicional se realizó para saber qué tan significativos son

los resultados del método FDRI en una situación parecida.

El método utilizado para el análisis estadístico fue *paired Student's t-test*, considerando como hipótesis nula la igualdad entre el resultado base y los resultados de FDRI. Los resultados del análisis se presentan con un error de $\alpha = 0.05$ al rechazar la hipótesis nula. Detalles del método *paired Student's t-test* y del proceso de validación estadística, se presentan en el apéndice D.

La tabla 6.9 muestra los resultados de la validación estadística realizada a los resultados del método FDRI. Entre paréntesis se muestra el *MAP* del promedio de los desempeños de las listas iniciales. La primera columna indica el método de fusión utilizado mientras que la columna **FS** muestra el *MAP* de la fusión sistemática. La columna **MR** muestra el *MAP* del mejor resultado del método FDRI y la columna **Conf** indica la medida de calidad utilizada y el número de listas incluidas en la fusión.

La tabla 6.9 muestra que, salvo en tres casos, el método FDRI logra obtener resultados estadísticamente signifcativos con un 95 % de confianza al compararlos con los de la fusión sistemática. De manera particular, en el conjunto **geoir08**, observamos que al fusionar con los métodos CombMNZ y Fuzzy Borda, se logra mejorar la efectividad global (*MAP*) pero la validación estadística muestra que esta mejora no es significativa. Este conjunto tiene dos características que derivan en este resultado: *a)* contiene muy pocos tópicos; *b)* los desempeños globales de sus listas iniciales tienen desempeños muy parecidos. La validación estadística es sensible a la primer característica, ya que tiene una tendencia inversamente proporcional entre el número de tópicos del conjunto y el aumento en la eficacia global de los resultados de recuperación. Por otro lado, los métodos CombMNZ y Fuzzy Borda utilizan la redundancia de los elementos en las listas, pero si éstas son muy parecidas la eficacia de estos métodos tiende a decrementar. Como se mostró en la tabla 6.5, las listas iniciales de este conjunto tienen la mayor intersección de elementos, lo cual explica la baja efectividad de los métodos CombMNZ y Fuzzy Borda en este conjunto.

Al considerar un escenario donde no se cuente con información de las listas, la validación estadística con respecto al promedio muestra que, salvo en un caso, no se logra superar la efectividad promedio de las listas. Este caso también se da en el conjunto **geoir08** con el método CombMNZ, que como se explicó anteriormente, decrementa su eficacia al fusionar listas muy parecidas. Estos resultados muestran que FDRI es una muy buena opción en escenarios donde exista un conjunto de listas de resultados para una misma petición, y se requiera ofrecer al usuario un sólo resultado

adhoc05			
(*efectividad promedio:* **0.250**)			
Método de fusión	**FS**	**MR**	**Conf.**
MaxRSV	0.231	$0.280^{\dagger\ddagger}$	Q_2 variable
CombMNZ	0.275	$0.300^{\dagger\ddagger}$	Q_4 2L
FuzzyBorda	0.267	$0.297^{\dagger\ddagger}$	Q_2 variable
geoir08			
(*efectividad promedio:* **0.231**)			
Método de fusión	**FS**	**MR**	**Conf.**
MaxRSV	0.180	$0.270^{\dagger\ddagger}$	Q_1 variable
CombMNZ	0.244	0.274	Q_4 3L
FuzzyBorda	0.265	$0.288^{\ddagger}$	Q_4 3L
image08			
(*efectividad promedio:* **0.238**)			
Método de fusión	**FS**	**MR**	**Conf.**
MaxRSV	0.251	$0.310^{\dagger\ddagger}$	Q_4 2L
CombMNZ	0.302	$0.340^{\dagger\ddagger}$	Q_4 3L
FuzzyBorda	0.316	$0.345^{\dagger\ddagger}$	Q_4 3L
robust08			
robust08 (*efectividad promedio:* **0.265**)			
Método de fusión	**FS**	**MR**	**Conf.**
MaxRSV	0.229	$0.320^{\dagger\ddagger}$	Q_2 variable
CombMNZ	0.341	$0.339^{\ddagger}$	Q_2 2L
FuzzyBorda	0.161	$0.303^{\dagger\ddagger}$	Q_1 variable

Tabla 6.9: Resultados de la validación estadística del mejor resultado obtenido por el método FDRI. †Significativo respecto a la fusión sistemática. ‡Significativo respecto al promedio.

de recuperación.

La tabla 6.9 muestra que las medidas Q_2 y Q_4 son las más efectivas al aplicar el método FDRI, sin embargo, son varias las configuraciones que ofrecen los mejores resultados lo cual es un claro indicio de que el uso de una u otra depende del conjunto de listas y del método de fusión utilizado.

Capítulo 7

Conclusiones y trabajo futuro

Este trabajo de investigación se propuso como objetivo el desarrollo de un método que ayudara a mejorar la efectividad de la Fusión de Datos sistemática. El problema existente en la fusión sistemática es que no siempre es conveniente fusionar todas las listas disponibles, y no siempre es conveniente fusionar las mejores listas, ya que suele suceder que para diferentes peticiones, un mismo sistema ofrezca desempeños variables. Por lo anterior se propuso una solución que conlleva una selección previa de las listas que van a fusionarse. El método resultante de esta investigación lleva el nombre de *Fusión Dinámica de Resultados de RI* (FDRI), ya que realiza un análisis previo de las listas de resultados de RI disponibles, y selecciona las más aptas para ser fusionadas con el fin de mejorar los resultados de la fusión. La parte dinámica del método se debe al hecho de que no son ni el mismo número ni las mismas listas las que se fusionan para distintas peticiones.

El método FDRI tiene las características de ser implementado en un enfoque no supervisado, ser independiente de los juicios de relevancia, ser independiente del funcionamiento interno de los sistemas y/o las estrategias de RI utilizados para obtener las listas de resultados, ser independiente de los métodos de fusión, y además de evitar la generación de todas las posibles fusiones de las listas disponibles. Lo anterior hace a nuestro método ideal para ser aplicado a cualquier contexto de RI en donde se generen múltiples resultados de recuperación para una misma petición y se desee utilizar Fusión de Datos.

Los análisis realizados en esta investigación y lo resultados experimentales obtenidos, nos permiten hacer las siguientes conclusiones:

- La Fusión de Datos permite tomar diferentes resultados de recuperación y generar una nueva lista de resultados, sin embargo es muy sensible a las listas con

malos resultados, las cuales hacen que la lista resultante de la fusión tenga una eficacia baja.

- Una selección previa de las listas que serán incluídas en el proceso de Fusión de Datos, en la cual se logre descartar a las listas con resultados de recuperación pobres, puede ayudar de manera efectiva a mejorar el resultado final de la fusión. El estudio manual realizado en esta investigación mostró que un proceso adecuado de selección, puede ofrecer un aumento sustancial en la efectividad de la fusión.

- Las características de redundancia y posicionamiento de los elementos recuperados mostraron ser efectivas para determinar la calidad de las listas de resultados. Las medidas de calidad basadas en estas dos características permitieron una discriminación adecuada de las listas que fueron incluídas en el proceso de fusión, lo cual permitió mejorar los resultados de la fusión sistemática.

- En un escenario donde se cuente con diferentes listas de resultados para una misma petición, pueden suscitarse diferentes casos: *a)* listas con resultados muy distintos; *b)* listas con resultados muy parecidos (redundantes); *c)* listas con buenos resultados; *d)* listas con malos resultados; *e)* combinaciones de las anteriores. Con esto en mente, podemos concluir que un método de selección de listas basado en redundancia y posicionamiento será ineficaz cuando: *i)* las listas de resultados contengan resultados muy distintos y una baja eficacia; *ii)* las listas de resultados contengan resultados muy parecidos y una alta eficacia. En particular, el método de selección desarrollado en esta investigación, tiene estas limitantes.

- Un factor muy importante en el proceso de RI es el soporte que la colección de búsqueda tenga para la necesidad de información de un usuario. Los conjuntos de peticiones considerados en esta investigación abarcan ambos casos: un buen soporte y un mal soporte. Este factor influyó directamente en el proceso de fusión, ya que con un buen soporte de peticiones se tuvo más oportunidad de contar con listas de resultados con una buena redundancia de los elementos relevantes. Por el contrario, con un mal soporte de peticiones, los elementos redundantes fueron en su mayoría no relevantes, lo cual influyó negativamente en los resultados de los métodos de fusión.

- Los métodos considerados en esta investigación fueron representativos de los métodos de fusión existentes, ya que se incluyó un método basado sólo en posicionamiento (MaxRSV), uno basado en posicionamiento y redundancia (CombMNZ), y uno basado en redundancia y el valor de confianza dado por el sistema de RI (Fuzzy Borda). Nuestros experimentos mostraron que los resultados del método basado sólo en posicionamiento, puede ser mejorado en gran medida al aplicar el método de selección de listas previa a la fusión. Por otro lado, nuestro método de selección de listas también fue efectivo para mejorar la efectividad de los otros dos métodos de fusión más sofisticados, por lo que podemos concluir que FDRI es independiente del método de Fusión de Datos utilizado.

- En el proceso de Fusión de Datos, conforme se incluyen más listas los elementos no relevantes aumentan considerablemente más que los elementos relevantes. Esta situación decrementa la efectividad de los métodos de fusión. Los experimentos con FDRI considerando un número de listas fijo son consistentes con esta observación, ya que puede notarse una tendencia donde la fusión de pocas listas obtiene buenos resultados, pero su efectividad decrece conforme se incluyen más listas en la fusión.

- Si bien la inclusión de un número mayor de listas contribuye al decremento de la efectividad de la Fusión de Datos, no hay un número exacto de listas que deben incluirse para obtener el mejor resultado. Nuestros experimentos mostraron que fusionar 2 o 3 listas ofrece los mejores resultados en la mayoría de los casos, esto puede no aplicarse a otro conjunto de listas de resultados. Lo anterior motivó experimentar la inclusión de un número variable de listas. Esta nueva estrategia logró mejorar los resultados de los métodos de fusión MaxRSV y Fuzzy Borda, los cuales tienen una base más fuerte en el posicionamiento. El método CombMNZ, basado más en redundancia, no se benefició con esta estrategia.

- El estudio de la efectividad particular que obtuvieron FDRI y la fusión sistemática para cada petición, mostró que el método FDRI logra que más peticiones obtengan resultados que superen a la mejor lista inicial. Sin embargo, en algunos casos no se logró superar la efectividad global de la fusión sistemática. Esta observación sugiere una situación donde se debe decidir si un resultado

global mayor es mejor que muchas mejoras de efectividad individuales. Esta decisión se deja al usuario o a la aplicación que utilice los resultados generados por nuestro método.

- De acuerdo a los resultados experimentales obtenidos y a las conclusiones anteriores, el método FDRI es efectivo para mejorar los resultados de la fusión sistemática en los conjuntos de datos considerados y con los métodos de fusión utilizados.

- En general, en un escenario de RI donde se cuente con múltiples resultados de recuperación para una misma petición, y se quiera aplicar Fusión de Datos para entregar un sólo resultados al usuario, el método FDRI es una mejor opción que la Fusión de Datos sistemática. Debido a sus resultados consistentes, la medida de calidad que se recomienda utilizar es Q_4. De acuerdo a los resultados obtenidos, se recomienda probar el método considerando las 3 mejores listas, o también el enfoque de listas variables.

7.1. Aportaciones

Las conclusiones anteriores nos permiten determinar las aportaciones de este trabajo de investigación, las cuales se listan a continuación:

1. Un conjunto de medidas que calculan la calidad de las listas de resultados de acuerdo a la redundancia y el posicionamiento de los elementos que contienen.

2. Un método de selección de listas de resultados de recuperación previa al proceso de Fusión de Datos que permite mejorar la efectividad del método de fusión a utilizar (FDRI), además de ser directamente aplicable un proceso de RI donde se tienen múltiples resultados de recuperación para una misma petición.

3. Un método que permite seleccionar para cada petición, un subconjunto de listas particular, de acuerdo a las características de redundancia y posicionamiento de los elementos en las listas de resultados.

4. Un método que tiene las características de ser implementado en un enfoque no supervisado, ser independiente de los juicios de relevancia, ser independiente del funcionamiento interno de los sistemas y/o las estrategias de RI utilizados para

obtener las listas de resultados, ser independiente de los métodos de fusión, y además de evitar la generación de todas las posibles fusiones de las listas disponibles.

5. Un análisis del poder de discriminación que tienen las características de redundancia y posicionamiento de los elementos en las listas, para determinar el subconjunto de listas que ofrecen una mejora en la efectividad de la fusión.

7.2. Trabajo futuro

Aún cuando los resultados obtenidos por el método FDRI son alentadores, existen algunas características que pueden ser incluidas y/o mejoradas. A continuación se listan estas características, las cuales conforman el trabajo futuro que se desprende de esta investigación:

- Debido a que algunas de las medidas de calidad propuestas se basan en la intersección de las listas de resultados, puede suceder que la intersección sea vacía, en cuyo caso todas las listas tendrían una calificación de 0,0. En este caso, el método eligiría a las listas en el orden que fueron introducidas al sistema. Por esta situación, deben buscarse diferentes formas de conformar el conjunto de elementos que ayudarán a las medidas de calidad a calificar a las listas. Dos posibles formas de hacerlo son las siguientes:

  1. **Fusión de datos redundante**. Este método se basa en el trabajo de Nuray [47], en el cual se utilizan los primeros n elementos resultantes de un proceso de fusión como elementos *pseudo-relevantes* que ayudarán a calificar a los sistemas de RI. En nuestro caso, esos n elementos podrían sustituir al conjunto intersección.
  2. **Características intrínsecas**. La formulación de nuevas características basadas en la similitud de la petición con el contenido de los elementos recuperados, podría ayudar a determinar un nuevo conjunto de elementos que sustituya al conjunto intersección. Para lo anterior podrían utilizarse técnicas de modelado de lenguaje.

- De acuerdo a los resultados obtenidos, no siempre una misma medida de calidad logra el mejor resultados para todos los conjuntos de listas. Una extensión a este

trabajo sería un análisis de las características que permiten a dichas medidas desempeñarse mejor en ciertos conjuntos, con lo cual se tendría la opción de variar también la medida de calidad en el proceso de selección de listas.

- Nuestros resultados muestran también que no siempre un método de fusión obtiene el mejor resultados para todos los conjuntos de listas. Por ejemplo, para los conjuntos **adhoc05** y **robust08** el método más adecuado es CombMNZ, mientras que para **geoir08** e **image08** lo es Fuzzy Borda. El poder variar el método de fusión, de acuerdo a características de las listas seleccionadas, permitiría al método obtener mejores resultados.

- El proceso de discriminación de listas basado en las diferencias de las medidas de calidad de las listas de resultados, permitió mejorar los resultados de la inclusión de un número fijo de listas en varios de los casos considerados. Sin embargo, una mejor discriminación debería ser capaz de mejorar todos los casos considerados. En este caso, se propone un nuevo esquema de discriminación basado en los valores mínimos y máximos que las medidas de calidad pueden obtener (apéndice C). Si el valor de calidad obtenido por una lista es muy bajo, comparado con el máximo valor que la medida puede obtener, tal vez sería mejor no incluir dicha lista.

Cada uno de los puntos anteriores podría aumentar la capacidad del método FDRI para mejorar la efectividad de diferentes métodos y estrategias de Fusión de Datos.

Apéndice

Publicaciones derivadas de la investigación

1. Júarez-González Antonio, Montes-y-Gómez Manuel, Villaseñor-Pineda Luis, and Ortiz-Arroyo Daniel. On the selection of the Best Retrieval Result per Query: An Alternative Approach to Data Fusion. In The Eight International Conference on Flexible Query Answering Systems (FQAS 2009), volume 5022/2009 of Lecture Notes in Artificial Intelligence, pages $111-121$, Roskilde, Denmark, October 2009. Springer.

2. Juárez-González Antonio, Montes-y-Gómez Manuel, Villaseñor-Pineda Luis, Pinto-Avendaño David, and Pérez-Coutiño Manuel. Selecting the N-top Retrieval Result Lists for an Effective Data Fusion. In The 11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2010), volume 6008/2010 of Lecture Notes in Computer Science, pages $580-589$, Romania, March 2010. Springer

En colaboración:

1. Hernández-Gracidas Carlos, Juárez-González Antonio, Sucar L. Enrique, Montes-y-Gómez Manuel, and Villaseñor-Pineda Luis. Data Fusion and Label Weighting for Image Retrieval based on Spatio-Conceptual Information. In The 9th international conference on Adaptivity, Personalization and Fusion of Heterogeneous Information, RIAO-2010. Paris, France, April 2010.

Apéndice

Tabla de acrónimos

AA	Aprendizaje Automático
AP	Average Precision
BR	Búsqueda de Respuestas
CLEF	Cross Language Evaluation Forum
EI	Extracción de Información
F	medida F
FD	Fusión de Datos
FDRI	Fusión Dinámica de Resultados de RI
FDS	Fusión de Datos Sistemática
GR	Generación de Resúmenes
idf	inverse document frequency
MAP	Mean Average Precision
ML	Mejor Lista de resultados
P	Precisión
Q	Medida de calidad
R	Recuerdo (recall)
r-prec	medida de precisión a r documentos
RI	Recuperación de Información
RSV	Retrieval Score Value
SRIG	Sistema de Recuperación de Información Geográfico
SRI	Sistema de Recuperación de Información
TAT	Tratamiento Automático de Texto
tf	term frequency
TREC	Text REtrieval Conference
WWW	World Wide Web

Apéndice

Valores máximos y mínimos de las medidas de calidad

Las medidas de calidad descritas en la sección anterior, actuarán sobre un conjunto de m listas de resultados de recuperación para calcular la calidad de éstas de acuerdo a la posición de los elementos más redundantes (intersección total). Es importante conocer los valores máximos y mínimos que nuestras medidas de calidad pueden tomar, con el fin de conocer qué tan buena es la calidad de las listas evaluadas con respecto a los límites teóricos de las medidas. A continuación presentamos el análisis particular de cada medida.

- Q_1. Esta medida suma las apariciones de los elementos de una lista en todo el conjunto de listas de resultados. Considerando cada lista como un conjunto de n elementos, tenemos dos casos:

  1. *Conjuntos disjuntos*. En este caso los elementos aparecen sólo una vez en el conjunto de listas (todos son elementos únicos), por lo que el valor máximo que puede obtener Q_1 es n. Más aún, éste es el caso en que la medida alcanza su valor mínimo, esto es:

  $$min(Q_1) = n$$

  2. *Conjuntos iguales*. En este caso los elementos aparecen m veces en el conjunto de listas, siendo m la cardinalidad del conjunto de listas L. Más aún, éste es el caso en que la medida de calidad Q_1 alcanzaría su máximo valor, esto es:

  $$max(Q_1) = m * n$$

Por tanto, podemos definir las cotas mínima y máxima para los valores que pueden obtenerse mediante la medida de calidad Q_1:

$$n \leq Q_1 \leq m * n \tag{C.0.1}$$

donde n es el tamaño de las listas y m es el número de listas consideradas.

- Q_2. De acuerdo a la definición de esta medida (ecuación 5.1.3), los cálculos se basan en la intersección total de las listas en L, por lo que en una situación donde listas fueran conjuntos disjuntos de elementos, la intersección sería vacía y por tanto el valor de esta medida sería 0, esto es:

$$min(Q_2) = 0$$

Por otro lado, cuando la intersección no es disjunta tenemos un conjunto de elementos y sus posiciones asociadas dentro de las listas de L. En este escenario, la situación que ofrecería el valor más alto sería cuando las listas contuvieran a los mismos elementos. En este caso, si las listas fueran de n elementos, tendríamos la sumatoria de n posiciones inversas, esto es:

$$\sum_{i=1}^{n} \left(\frac{1}{i}\right) = \frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{n}$$

La sumatoria anterior describe una *serie armónica parcial*, la cual puede ser calculada de la siguiente manera:

$$\sum_{i=1}^{n} \left(\frac{1}{i}\right) \approx ln(n) + r + \frac{1}{2n} + \frac{1}{12n^2} + \frac{1}{120n^4} - \varepsilon$$

donde $r = 0{,}57721566$ es la constante de Euler-Mascheroni y $0 < \varepsilon < \frac{1}{252n^6}$.

Con lo anterior, podemos definir las cotas mínima y máxima para los valores que pueden obtenerse mediante la medida de calidad Q_2:

$$0 \leq Q_2 \lesssim ln(n) + r + \frac{1}{2n} + \frac{1}{12n^2} + \frac{1}{120n^4} - \varepsilon \tag{C.0.2}$$

donde n es el tamaño de las listas consideradas.

- Q_3. Esta medida también se basa en la intersección total de las listas en L, por lo que el valor mínimo se obtiene cuando la intersección es vacía, esto es:

$$min(Q_3) = 0$$

Por otro lado, el valor máximo se obtiene cuando las listas son conjuntos de elementos idénticos, con lo cual tenemos el inverso de la suma de todas las posiciones:

$$\frac{1}{1+2+3+\cdots+(n-1)+n}$$

El denominador se sustituye por la fórmula de la suma de los primeros n números consecutivos, con lo cual obtenemos:

$$\frac{1}{\left(\frac{n(n+1)}{2}\right)} = \frac{2}{n(n+1)}$$

Con lo anterior, el valor máximo para la medida Q_3 es:

$$max(Q_3) = \frac{2}{n(n+1)}$$

De esta manera definimos las cotas mínima y máxima para los valores de Q_3:

$$0 \leq Q_3 \leq \frac{2}{n(n+1)} \tag{C.0.3}$$

donde n es el tamaño de las listas consideradas.

- Q_4. Como en los dos casos anteriores, esta característica se basa en la intersección total de L, por lo que el valor más bajo que puede obtenerse es cuando se sucita una intersección vacía, esto es:

$$min(Q_4) = 0$$

Como en el caso de Q_2 y Q_3, el máximo valor posible se obtiene con listas de resultados idénticas. Para calcular este valor, calculemos los valores de las primeras posiciones, considerando un tamaño de n elmentos para las listas de resultados. De esta manera:

- Para el primer elemento de la lista:

$$1 - \frac{ln(1)}{ln(n)}$$

- Para el primer y segundo elemento de la lista:

$$1 - \frac{ln(1)}{ln(n)} + 1 - \frac{ln(2)}{ln(n)} = 2 - \left(\frac{ln(1)}{ln(n)} + \frac{ln(2)}{ln(n)}\right) =$$

$$2 - \left(\frac{ln(1)+ln(2)}{ln(n)}\right) = 2 - \left(\frac{ln(1*2)}{ln(n)}\right) =$$

$$2 - \frac{ln(2)}{ln(n)}$$

- Para el primero, segundo y tercer elemento de la lista:

$$1 - \frac{ln(1)}{ln(n)} + 1 - \frac{ln(2)}{ln(n)} + 1 - \frac{ln(3)}{ln(n)} = 3 - \left(\frac{ln(1)}{ln(n)} + \frac{ln(2)}{ln(n)} + \frac{ln(3)}{ln(n)}\right) =$$

$$3 - \left(\frac{ln(1)+ln(2)+ln(3)}{ln(n)}\right) = 3 - \left(\frac{ln(1*2*3)}{ln(n)}\right) =$$

$$3 - \frac{ln(6)}{ln(n)}$$

- Para el primero, segundo, tercer y cuarto elemento de la lista:

$$1 - \frac{ln(1)}{ln(n)} + 1 - \frac{ln(2)}{ln(n)} + 1 - \frac{ln(3)}{ln(n)} + 1 - \frac{ln(4)}{ln(n)} =$$

$$4 - \left(\frac{ln(1)}{ln(n)} + \frac{ln(2)}{ln(n)} + \frac{ln(3)}{ln(n)} + \frac{ln(4)}{ln(n)}\right) =$$

$$4 - \left(\frac{ln(1)+ln(2)+ln(3)+ln(4)}{ln(n)}\right) = 4 - \left(\frac{ln(1*2*3*4)}{ln(n)}\right) =$$

$$4 - \frac{ln(24)}{ln(n)}$$

En este punto podemos ver una tendencia al aumentar el número de elementos. Así para los i primeros elementos, el valor de Q_4 es:

$$i - \left(\frac{ln(1*2*\cdots*(i-1)*i)}{ln(n)}\right) = i - \frac{ln(i!)}{ln(n)}, \quad 1 \leq i \leq n$$

Con lo anterior, podemos definir las cotas mínima y máxima para los valores que pueden obtenerse mediante la medida de calidad Q_4:

$$0 \leq Q_4 \leq n - \frac{ln(i!)}{ln(n)}, \quad 1 \leq i \leq n \tag{C.0.4}$$

donde n es el tamaño de las listas consideradas.

Apéndice

Método de validación estadística *paired Student's t-test*

Cuando se requiere conocer si la mejora en el desempeño de un sistema no es por casualidad, la forma más adecuada de hacerlo es utilizando *métodos de validación estadística*. Existen diferentes métodos de validación estadística cuyo propósito es ofrecer una herramienta para rechazar una *hipótesis nula*. Esta hipótesis nula normalmente es que *los resultados que se comparan son iguales*. De esta manera, dependiendo del método de validación, se cuentan con tablas de valores que delimitan los umbrales de error con los cuales se rechaza la hipótesis nula.

En el área de Recuperación de Información, dado un conjunto de peticiones y dos resultados de recuperación para cada petción, uno de los métodos utilizados para determinar cuándo un resultado es significativamente mejor que el otro, es el llamado *paired Student's t-test*.

La aplicación de este método se requiere que las observaciones para las dos se obtengan en pares, y que dichas observaciones se realicen en condiciones idénticas. Dado que la Recuperación de Información para una misma petición en la misma colección tiene las mismas condiciones, este método es directamente aplicable a los resultados de recuperación.

Sean dos listas de resultados de recuperación $l_1 =< s_{1,1}, \ldots, s_{1,n} >$ y $l_2 =< s_{2,1}, \ldots d_{s,n} >$, donde cada $s_{i,j}$ es el valor de similitud para cada elemento recuperado dado por el SRI con el que se obtuvo la lista. Los pasos a seguir para aplicar el método son los siguientes:

1. Obtener el promedio de las observaciones para cada lista:

$$\bar{x}_1 = \frac{\sum_{i=1}^{n} s_{1,i}}{n} \quad \text{y} \quad \bar{x}_2 = \frac{\sum_{i=1}^{n} s_{2,i}}{n}$$

2. Calcular la diferencia entre observaciones:

$$r_i = s_{1,i} - s_{2,i}.$$

3. Obtener el promedio de las diferencias:

$$\bar{r} = \frac{\sum_{i=1}^{n} r_i}{n}$$

4. Calcular la varianza de las diferencias:

$$s^2 = \sum_{i=1}^{n} \frac{(s_{1,i} - \bar{r})^2}{n-1}$$

5. Obtener el valor t que será utilizado para compararlo con los umbrales de error:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s/n^2}$$

El valor t obtenido nos servirá para determinar si los resultados en la lista l_1 tienen una mejora estadísticamente significativa sobre los resultados en la lista l_2. Dependiendo del umbral que sobrepase t, el valor α asociado nos dirá el error que el método tiene al rechazar la hipótesis de que los resultados de l_1 y l_2 son iguales (hipótesis nula). Los umbrales de error para este método se muestran en la figura D.1.

Si el valor t es mayor al umbral marcado para $n-1$ observaciones, entonces la mejora en los resultados de la lista l_1 son estadísticamente significativos con el error marcado para el umbral de α sobrepasado por t. Por ejemplo, si hay 10 observaciones y t =2.1, entonces la mejora del desempeño de l_1 sobre l_2 es estadísticamente significativa con un error de $\alpha = 0{,}05$.

Más acerca de este método de validación estadística, así como métodos de validación adicionales pueden consultarse en [30]

	Level of significance α				
ν	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
∞	1.282	1.645	1.960	2.326	2.576

Figura D.1: Valores t críticos.

Bibliografía

[1] Spoerri A. Using the structure of overlap between search results to rank retrieval systems without relevance judgments. *Information Procesing and Management*, 4(43):1059–1070, 2007.

[2] Woodley Alan, Lu Chengye, Sahama Tony, King John, and Geva Shlomo. Queensland university of technology at trec 2005. *In Proceedings of the Fourteenth Text REtrieval Conference*, 2005.

[3] Zazo Angel F., Figuerola Carlos G., and Alonso-Berrocal José L. Reina at clef 2007 robust task. *In Proceedings of the Cross Language Evaluation Forum (CLEF)*, 2007.

[4] Juárez-González Antonio, Montes-y-Gómez Manuel, Villaseñor-Pineda Luis, and Ortiz-Arroyo Daniel. On the selection of the best retrieval result per query: An alternative approach to data fusion. In *The Eight International Conference on Flexible Query Answering Systems (FQAS 2009)*, volume 5022/2009 of *Lecture Notes in Artificial Intelligence*, pages 111–121, Roskilde, Denmark, October 2009. Springer.

[5] Juárez-González Antonio, Montes-y-Gómez Manuel, Villaseñor-Pineda Luis, Pinto-Avendaño David, and Pérez-Coutiño Manuel. Selecting the n-top re-trieval result lists for an effective data fusion. In *The 11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2010)*, volume 6008/2010 of *Lecture Notes in Computer Science*, pages 580–589, Romania, March 2010. Springer.

[6] Bartell B. T., Cottrell G. W., and Belew R. K. Automatic combination of multiple ranked retrieval systems. In *Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1994.

[7] Heuwing Ben and Mandl Thomas. Robust retrieval experiments at the university of hildesheim. *In Proceedings of the Cross Language Evaluation Forum (CLEF)*, 2007.

[8] Peters C. What happened in clef 2008 introduction to the working notes. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[9] Vogt C. C. and Cottrell G. W. Predicting the performance of linearly combined ir systems. In *21st ACM-SIGIR International Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 1998.

[10] Vogt C. C., Cottrell G. W., Belew R. K., and Bartell B. T. Using relevance to train a linear mixture of experts. In *The Fifth Text REtrieval Conference (TREC-5)*, NIST Special Publication, Gaithersberg, MD. National Institute of Standards and Technology, 1997.

[11] Manning Christoper D., Raghavan Prabhakar, and Schütze Hinrich. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England, 2009.

[12] Vogt Christopher and Garrison Cottrell. Fusion via a linear combination of scores. *Information Retrieval*, 1(3):151–173, 1999.

[13] Kompaoré D., Mothe J., A. Baccini, and S. Dejean. Query clustering to decide the best system to use. In *The RIAO 2007, 8th International Conference*, 2007.

[14] Lillis D., Toolan F., Collier R., and Dunnion J. Probfuse: A probabilistic approach to data fusion. In *The International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 139–146, 2006.

[15] Grossman D. A. and Frieder O. *Information Retrieval, Algorithms and Heuristics*. Springer, second edition, 2004.

[16] Hsu D. F. and Taksa I. Comparing rank and score combination methods for data fusion in information retrieval. *Information Retrieval*, 8(3):449–480, 2005.

[17] Roussinov Dimitri, Chau Michael, Filatova Elena, and Robles-Flores José Antonio. Building on redundancy: Factoid question answering, robust retrieval and the "other". *In Proceedings of the Fourteenth Text REtrieval Conference*, 2005.

[18] Metzler Donald, Diaz Fernando, Strohman Trevor, and Croft W. B. Umass robust 2005: Using mixtures of relevance models for query expansion. *In Proceedings of the Fourteenth Text REtrieval Conference*, 2005.

[19] Agirre E., DiÑunzio G. M., Ferro N., Mandl T., and Peters C. Clef 2008: Ad hoc track overview. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[20] Villatoro-Tello E., Montes-y-Gómez M., and Villaseñor-Pineda L. Inaoe at geoclef 2008: A ranking approach based on sample documents. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[21] Fox E. A. and Shaw J. A. Combination of multiple searches. In *The Second Text REtrieval Conference, TREC-2*, 1994.

[22] Vorhees E. M. Overview of trec 2007. In *The Sixteenth Text Retrieval Conference (TREC)*, 2007.

[23] Yom-Tov Elad, Fine Shai, Carmel David, Darlow Adam, and Amitay Einat. Improving document retrieval according to prediction of query dificult. *In Proceedings of the Thirteenth Text REtrieval Conference*, 2004.

[24] Gantz F., Reinsel D., Chute C., Schlichting W., McArtur J., Minton S., Xheneti I., Toncheva A., and Manfrediz A. A forecast of worldwide information growth through 2010. *IDC - The Expanding Digital Universe*, March 2007.

[25] Picouto F., Lorente I., García-Moran J. P., and Ramos A. *Hacking y Seguridad en Internet*. Alfaomega, 2008.

[26] Martínez-Santiago Fernando, Montejo-Ráez Arturo, and García-Cumbreras Miguel A. Sinai at clef ad-hoc robust track 2007: applying google search engine for robust cross-lingual retrieval. *In Proceedings of the Cross Language Evaluation Forum (CLEF)*, 2007.

[27] Hubert G. and Mothe J. Relevance feedback as an indicator to select the best search engine - evaluation on trec data. In *the Ninth International Conference on Enterprise Information Systems, ICEIS*, 2007.

[28] Lebanon G. and Lafferty J. Cranking: Combining rankings using conditional probability models on permutations. In *The Nineteenth International Conference on Machine Learning*, 2002.

[29] Salton G. *Automatic Information Organization and Retrieval*. McGraw-Hill, New York, 1968.

[30] Kanji G. K. *100 Statistical Tests*. Sage Publications, third edition, 2006.

[31] Ed Greengrass. Information retrieval: A survey, November 2000.

[32] Ding Guodong, Wang Bin, and Bai Shuo. Cas-ict at trec 2005 robust track: Using query expansion and rankfusion to improve effectiveness and robustness of ad hoc information retrieval. *In Proceedings of the Fourteenth Text REtrieval Conference*, 2005.

[33] Escalante Hugo Jair, Hérnadez Carlos A., Sucar Luis Enrique, and Montes-y-Gómez Manuel. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proceeding of the 1st ACM international conference on Multimedia information retrieval*, MIR '08, pages 172–179. ACM, 2008.

[34] Aslam J. A. and Montague M. Models for metasearch. In *The 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 276–284, 2001.

[35] Lee J. H. Analyses of multiple evidence combination. In *20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1997.

[36] Perea J. M., Ureña L. A., Buscaldi D., and Rosso P. Textmess at geoclef 2008: Result merging with fuzzy borda ranking. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[37] Ng K. B. and Kantor P. Predicting the effectiveness of naive data fusion on the basis of system characteristics. *Journal of American Society for Information Science*, (51):1177–1189, 2000.

[38] Yang Kiduk, YuÑing, George Nicholas, Loehrlen Aaron, McCaulay David, Zhang Hui, Akram Shahrier, Mei Jue, and Record Ivan. Widit in trec-2005 hard, robust, and spam tracks. *In Proceedings of the Fourteenth Text REtrieval Conference*, 2005.

[39] Kwok K.L., Grunfeld L., Dinstl N., and Deng P. Trec2005 robust track experiments using pircs. *In Proceedings of the Fourteenth Text REtrieval Conference*, 2005.

[40] Farah M. and Vanderpooten D. An outranking approach for rank aggregation in information retrieval. In *The 30th Annual International ACM SIGIR Conference*, pages 591–598, 2007.

[41] Montague M. and Aslam J. A. Condorcet fusion for improved retrieval. In *The 11th International Conference on Information Knowledge and Management (CIKM), ACM*, pages 538–548, 2002.

[42] Berry Michael W. and Browne Murray. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. Software, Environments, and Tools. SIAM, second edition edition, April 2005.

[43] Mark Montague and Javed A. Aslam. Relevance score normalization for metasearch. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, New York, NY, USA, 2001.

[44] Belkin N. J., Kantor P., Fox E. A., and Shaw J. A. Combining the evidence of multiple query representations for information retrieval. *Information Processing and Management*, 31(3):431–448, 1995.

[45] Gopalan N. P. and Batri K. Adaptive selection of top-m retrieval strategies for data fusion in information retrieval. *International Journal of Soft Computing*, 1(2):11–16, 2007.

[46] Kantor P. B. Decision level data fusion for routing of documents in the trec3 context: A best case analysis of worst case results. In *Third Text REtrieval Conference (TREC-3)*, NIST Special Publication, Gaithersberg, MD. National Institute of Standards and Technology, 1997.

[47] Nuray R. and Can F. Automatic ranking of information retrieval systems using data fusion. *Information Processing and Management*, 3(42):595–614, 2006.

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

[49] Wu S. and Crestani F. Data fusion with estimated weights. In *Eleventh International Conference on Information and Knowledhe Management*, McLean, USA, 2001.

[50] Wu S. and Crestani F. Methods for ranking information retrieval systems without relevance judgments. In *The 2003 ACM Symposium on Applied Computing*, pages 811–816, 2004.

[51] Liu Shuang and Yu Clement. Uic at trec2005: Robust track. *In Proceedings of the Fourteenth Text REtrieval Conference*, 2005.

[52] Arni T., Clough P., Snderson M., and Grubinger M. Overview of the imagecleffphoto 2008 photographic retrieval task. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[53] Diamond T. and Liddy E. D. Dynamic data fusion. In *Workshop of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, Maryland, 1998.

[54] Mandl T., Carvalho P., Gey F., Larson R., Santos D., and Womser-Hacker C. Geoclef 2008: the clef 2008 cross-language geographic information retrieval track overview. In *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, 2008.

[55] Shengli Wu and Sally Mcclean. Performance prediction of data fusion for information retrieval. *Information Processing and Management*, 42(4):899–915, 2006.

[56] Bernstein Yaniv, Billerbeck Bodo, Garcia Steven, Lester Nicholas, Scholer Falk, and Zobel Justin. Rmit university at trec 2005: Terabyte and robust track. *In Proceedings of the Fourteenth Text REtrieval Conference*, 2005.