Universidad Nacional de Educación a Distancia
Escuela Técnica Superior de Ingeniería Informática
Departamento de Lenguajes y Sistemas Informáticos

Predicción del rendimiento de consultas basado en rankings de documentos y nuevo marco de evaluación.

Tesis Doctoral

Joaquín Pérez Iglesias
Ingeniero Informático

2012

Universidad Nacional de Educación a Distancia
Escuela Técnica Superior de Ingeniería Informática
Departamento de Lenguajes y Sistemas Informáticos

Predicción del rendimiento de consultas basado en rankings de documentos y nuevo marco de evaluación.

Tesis Doctoral

Joaquín Pérez Iglesias

Ingeniero en Informática por la Universidad Rey Juan Carlos de Madrid.

Director:

Lourdes Araujo Serna

Profesor Titular de Universidad del Departamento de Lenguajes y Sistemas Informáticos
de la Universidad Nacional de Educación a Distancia

A mi familia.

Agradecimientos

Esta tesis no podría considerarse completa sin agradecer a todas aquellas personas que, de alguna forma, han sido una parte fundamental en este largo camino.

En primer lugar, por supuesto, debo agradecer el trabajo realizado a mi directora Lourdes, por su incansable tenacidad, por su infinita paciencia y por sus incontables visitas, y cuando digo incontables digo bien. Sorprendentemente casi siempre sabía donde encontrarme, aunque nunca se lo puse fácil, y ha sabido apretarme lo suficiente para que pudiera llegar a escribir estas últimas líneas de mi tesis. ¡Muchas gracias Lourdes!

Por supuesto también debo agradecer a Víctor su participación en esta tesis. Siempre ha estado a mi lado en los buenos y malos momentos, afortunadamente creo que han sido más de los primeros. Por su apoyo desde los tiempos de Móstoles, pasando por el Carlota, el DEA en la Complutense, hasta llegar a esta última época en la UNED. Por estar siempre disponible y sobre todo porque es un placer trabajar con alguien con el que puedes reirte en cualquier situación. ¡Muchas gracias Fresno!

No podría olvidarme de José Ramón, quien de alguna forma me devolvió la pelota al meterme en esto de la recuperación de información. Aún no tengo claro si todo esto formaba parte de algún tipo de cruel venganza por aquella lejana historia en las escaleras del metro, pero aunque así fuera desde luego ha merecido la pena. ¡Muchas gracias Agüera!

Por supuesto a mi familia, que hasta donde me alcanza la memoria siempre ha apoyado totalmente mis decisiones, por otro lado algo bastante sorprendente porque algunas de ellas eran bastante peregrinas. ¡Muchas gracias familia!

A mis compañeros del departamento, a los cosanguineos Juaner, Alberto, Arkaitz, a los economistas Raquel, Anselmo, a los del pasillo Emilio, Álvaro, Damiano, Guille, Peinado ..., a todos. Siempre habéis estado disponibles para un profunda charla sobre primas de riesgo, diversos conflictos mundiales o bien comentar la última de Mourinho, pero sobre todo para aconsejarme y ayudarme en el desarrollo de esta tesis. ¡Muchas gracias compañeros!

También debo, y quiero, agradecer el apoyo de mis amigos durante estos últimos meses. Siempre disponibles para superar las distintas crisis que iban apareciendo, y siempre superadas favorablemente en el *Clover*. Mucho me

temo que probablemente en un futuro no muy lejano nos pasará factura. ¡Muchas gracias amigos!

En este espacio es imposible nombrar a toda la gente que de alguna manera han tenido algo que ver no solo en esta tesis, sino en lo que me he convertido, y aunque me gustaría no puedo nombrar a todos y así hacer que su nombre aparezca en estos agradecimientos. Por tanto, como agradecimiento a todos ellos y para no olvidarme de ninguno, solo me queda decir ¡Muchas gracias a todos!

Resumen

El tema principal en el que se encuadra este trabajo de tesis es el de la predicción de la calidad de consultas o *Query Performance Prediction (QPP)*. Este campo, que se engloba dentro del ámbito de la recuperación de información, ha intensificado su desarrollo de forma muy extensa durante los últimos años. El objetivo de estas técnicas de predicción es el de estimar la calidad del conjunto de documentos recuperados a partir de una consulta.

Durante los últimos años han ido apareciendo distintas propuestas que intentan explotar características de la consulta, de los documentos en la colección u otras fuentes externas con el objetivo de realizar las estimaciones.

Los distintos métodos de predicción se clasifican en dos grupos principales. Los métodos denominados *Post-Retrieval*, hacen uso de la lista de documentos recuperados a partir de una consulta y los denominados *Pre-Retrieval*, siendo más simples, realizan las estimaciones sin hacer uso de esta lista de resultados.

En esta tesis se presenta una nueva propuesta de predicción de tipo *Post-Retrieval*, basada exclusivamente en los valores de relevancia que asigna una función de ranking a los documentos que se recuperan a partir de una consulta. La hipótesis principal que está detrás de esta nueva propuesta, es el hecho de que un alto grado de dispersión entre los valores de relevancia implica una mayor calidad de la respuesta. Esta hipótesis surge a partir de ciertas características inherentes a un gran conjunto de funciones de ranking, que generan valores que pueden ser considerados como estimaciones cuantitativas de la probabilidad de que un documento sea relevante dada una consulta. Para que estas funciones de ranking muestren un grado de rendimiento aceptable, deben tener la capacidad de discriminar documentos relevantes de no relevantes a partir de los valores que asignan. Así, si en la lista de documentos devueltos se observa un alto grado de dispersión entre los valores de relevancia, se puede inferir que la función de ranking ha sido efectiva discriminando documentos relevantes de no relevantes.

Los resultados que se obtienen con esta nueva aproximación para predecir la calidad de una consulta, están en el entorno de aquellos considerados como más efectivos dentro del campo de la predicción, como *Clarity Score*. Además, esta propuesta cuenta como principal ventaja con el hecho de realizar estimaciones sin necesidad de aplicar técnicas complejas, ya que las

predicciones se calculan en base a la desviación estándar.

A continuación, el contenido de esta tesis se dirige de forma más específica al marco de evaluación que se emplea en la actualidad para medir el rendimiento de los métodos de predicción. Con anterioridad a este trabajo, algunos autores han destacado algunas de las limitaciones de este marco basado únicamente en la correlación que aparece entre las predicciones y la calidad de la respuesta medida con los juicios de relevancia.

Este marco de evaluación sufre las limitaciones típicas que presentan los distintos coeficientes de correlación. Además, de forma específica en el entorno de la predicción de consultas, al medir el rendimiento en base a la correlación se ignoran comportamientos específicos de los métodos de predicción. Así, se podría estar interesado, no en el rendimiento global que muestra un método de predicción, sino en su efectividad respecto a consultas de mayor o menor calidad.

Con el objetivo de evitar estas limitaciones, se propone un nuevo marco de evaluación, en el que el rendimiento de los distintos métodos se evalúa en base a su acierto, a la hora de clasificar las consultas en distintos grupos. Los distintos grupos de consultas se construyen automáticamente en base a la calidad que muestran según una medida de evaluación como podría ser la precisión media.

Finalmente, se analiza la utilidad de las predicciones en un entorno concreto como el de la expansión de consultas selectiva. Así, el objetivo es medir el rendimiento potencial que tendría un sistema de recuperación de información, en el caso de que contara con estimaciones sobre la calidad de las consultas que expresa el usuario. En base a estas estimaciones, un sistema podría decidir en qué casos la expansión proporcionaría un rendimiento superior a la consulta, y por tanto debiera realizarse.

A partir de este estudio, se comprueba experimentalmente que la medida de precisión media no es un buen estimador a la hora de realizar expansión selectiva, lo que hace inadecuada su predicción en un entorno como éste. Sin embargo, se corrobora que la predicción de una medida como *P@10* sería más adecuada para la expansión selectiva.

Abstract

This thesis is focused on the field of Query Performance Prediction (QPP). This subject, which is part of Information Retrieval research area, has received increasing attention in the last years. The main purpose of these prediction techniques is to estimate the quality of the document set retrieved when a query is posed to a search system.

Different proposals have been introduced to tackle this problem. They try to exploit specific properties from the query, the document collection or any other information source. This QPP techniques fall within two main approaches: Pre-Retrieval, where the ranking list returned by the search system is ignored and thus this list is not considered, and Post-Retrieval, where the returned documents are analysed to improve the quality of the final estimations.

This thesis introduces a new Post-Retrieval query performance prediction technique. The proposed method is based on the scores assigned by a ranking function to the returned documents. The main hypothesis behind this technique is that a high dispersion along the document scores could imply a high quality of the search system response. This idea raises from some of the ranking functions characteristics, since the computed relevance scores can be observed as quantitative estimations of the documents relevance probability.

The results obtained with this approach are similar to those obtained with other prediction methods considered accurate, such as Clarity Score. In addition, this approach has the advantage of performing the process without using complex techniques, since the predictions are computed by using the standard deviation.

Next, this thesis is focused on the current evaluation framework for prediction methods. Previously, some authors have emphasized some of the main disadvantages that appear with this evaluation framework based on the correlation found between the query predictions and the quality of the queries measured using the relevance judgements.

This evaluation framework shows some of the typical drawbacks due to an evaluation based on correlation. In addition, for the QPP case, measuring the performance of a prediction technique with a correlation coefficient ignores some of its specific properties, since the obtained correlation provides

only a measure of the global prediction accuracy. However, for some applications it could be interesting to evaluate the specific predictor performance for high or low quality queries.

In order to avoid the drawbacks mentioned above, a new evaluation framework is introduced in this thesis. This new approach measures the performance of a prediction method classifying different queries by their predicted quality.

Finally, the application of prediction methods to the selective query expansion scenario is also studied. The main purpose of this analysis is to measure the impact of applying these methods within a selective query expansion scenario. That is, the goal is to help a search system to decide in which cases the expansion of a query would improve the quality of the returned documents and thus whether the expansion process should be applied.

This analysis has experimentally proved the inadequacy of the average precision measure as an estimator to decide in which cases a query should be expanded. Besides, it is showed the suitability of predictions based on estimating the P@10 value instead of the average precision obtained by a query in this scenario.

Índice general

Índice de Figuras

Índice de Tablas

Most of the fundamental ideas of science are essentially simple, and may, as a rule, be expressed in a language comprehensible to everyone.

Albert Einstein.

Capítulo 1

Introducción

It is a very sad thing that nowadays there is so little useless information.

Oscar Wilde.

En la actualidad vivimos en un entorno digital en el que la tarea de buscar información en la Web se ha convertido en una actividad rutinaria, tanto en entornos profesionales como en otros más cercanos al ocio. La acción de buscar información en Internet utilizando los clásicos buscadores generalistas, como Google, Yahoo o Bing, es realizada por millones de usuarios diariamente, y en muchos de los casos dichas búsquedas proporcionan resultados, que en general pueden considerarse como suficientemente satisfactorios (ACSI. (2011)).

Una gran parte de la teoría aplicada al proceso de búsqueda en la Web, deposita sus bases en el campo de la Recuperación de Información o RI cuyos orígenes se remontan a la década de los 50. Hoy en día se puede considerar como precursora de los motores de búsqueda, tal y como los conocemos en la actualidad, la propuesta realizada por Vannevar Bush en 1945 (Bush (1945)) en su artículo *"As we may think"*, la cuál, en aquella época, se encontraba posiblemente más cercana a la ciencia ficción que a un futuro cercano. En este artículo, se describía un sistema automático con capacidad para almacenar una gran cantidad de información proveniente de diversas fuentes, que el autor denominó "memoria colectiva". A partir de este sistema sería posible recuperar de forma sencilla distintos elementos de conocimiento a petición del usuario.

Posteriormente, ya entrada la década de los 60, la Universidad de Cornell desarrolla el sistema *SMART* (Salton y Lesk (1965)), que se considera como uno de los ejemplos más significativos de lo que sería un Sistema de Recuperación de Información o SRI y que en cierta manera se acercaba a aquellas expectativas expresadas por Vannevar Bush.

1.1. Sistemas de recuperación de información (SRI)

En una primera aproximación se puede considerar que un SRI tiene como funcionalidad principal la capacidad de, a partir de una consulta interpre-

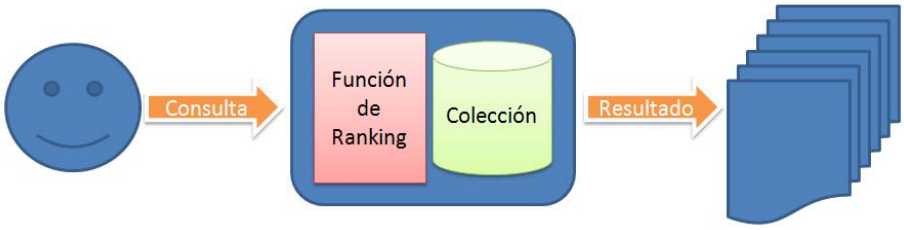

Figura 1.1: Componentes básicos que forman parte de un Sistema de Recuperación de Información.

table por dicho sistema, devolver un conjunto de unidades de información ordenadas según el grado de "utilidad" para dicho usuario. Por tanto el SRI deberá seleccionar entre el total de información que contenga, qué elementos o documentos son los más adecuadas dada una consulta. Es decir, para la tupla *usuario#consulta* el sistema debe devolver un conjunto de documentos ordenados por su "utilidad" para el usuario. La relación entre estos componentes aparece simplificada en la Figura 1.1

Aunque un SRI debe dar solución a un amplio rango de tareas tales como: la indexación de contenidos, la representación de estos contenidos o las interfaces de respuesta al usuario, la motivación principal de este trabajo gira en torno a mejorar la calidad de los resultados que se suministran al usuario. Esto es, la capacidad del SRI para devolver documentos relevantes para un *usuario#consulta*, o en palabras de Van Rijsbergen "... la capacidad para recuperar todos los documentos *relevantes* y simultáneamente recuperar el mínimo número posible de documentos *no relevantes*" (van Rijsbergen (1979), pág. 6).

Uno de los conceptos más importantes y que ha generado una mayor controversia dentro de la RI, es el concepto de relevancia[1]. En este trabajo se define un documento relevante como aquel que cumple las expectativas de un usuario, una vez que éste ha expresado una necesidad de información.

Formalmente, una función de ranking se puede definir como una función que representa la relación entre las distintas unidades de información *(O)*, y una necesidad de información *(IN)* expresada por el usuario en forma de consulta *(Q)*. Es importante destacar que la función de ranking debe incorporar la información implícita *(I)* que subyace de la necesidad de información. Esta información implícita, que no aparece reflejada en la consulta Q puede incluir por ejemplo las preferencias o el contexto del usuario, los conocimientos previos de éste o simplemente su estado de ánimo. Así, podemos definir una función de ranking $\Re$ como aquella que asigna valores de relevancia,

[1]Una discusión más amplia sobre el concepto de relevancia puede encontrarse en Borlund (2003) y en Mizzaro (1997).

basándose en la relación existente entre un conjunto de documentos y una necesidad de información:

$$\Re(O, IN) = \Re(O, (I, Q)) \tag{1.1}$$

A lo largo de los años distintos marcos teóricos han servido para modelar la relación a la que debe dar soporte una función de ranking. Así, aparecen modelos como el booleano, modelo de espacio vectorial, modelo probabilístico, modelos basados en lógica difusa o más recientemente distintos modelos inspirados en la teoría cuántica.

Aunque en general los distintos modelos de recuperación de información que han sido propuestos a lo largo de los años muestran un comportamiento muy robusto, la mayoría de ellos tienen como principal limitación la forma en la que el usuario expresa su necesidad de información. Así, si un usuario introduce una consulta eminentemente ambigua o la expresa usando un vocabulario que no se corresponde con el lenguaje con el que se representa dicho concepto en el conjunto del corpus indexado, dicha consulta, muy posiblemente, obtendrá como resultado un conjunto de documentos poco relevantes.

1.2. Expansión de consultas

Una aproximación clásica para aumentar el rendimiento de dichos modelos mejorando la representatividad de una consulta, parte de lo que se denomina refinamiento de la consulta o expansión de consultas. Dicha aproximación se basa en modificar la consulta original Q expresada por el usuario, añadiendo o eliminando términos, de forma que la nueva consulta Q' tenga una mayor capacidad para recuperar documentos relevantes. El enfoque más común para la selección de nuevos términos radica en su extracción a partir de una muestra de documentos que se consideren relevantes.

La selección del conjunto de documentos relevantes que sirve de ejemplo se obtiene generalmente utilizando dos aproximaciones distintas (Ruthven y Lalmas (2003)):

- La primera aproximación se denomina *Relevance Feedback (RF)* y exige que el usuario indique explícitamente los documentos que considera relevantes. Así, de forma interactiva y partiendo de los documentos devueltos a partir de la consulta original, se muestra al usuario un conjunto de documentos que el SRI considera relevantes. De este conjunto de documentos el usuario debe seleccionar aquellos que considera realmente relevantes de acuerdo a sus expectativas. Basándose en este subconjunto, el SRI generará una nueva consulta Q' que incluirá nuevos términos extraídos de los documentos seleccionados. La gran ventaja de este método es que es el mismo usuario el que indica

lo que considera relevante en ese instante y lo que no, por lo se está incorporando a la ecuación 1.1 la información implícita *(I)* que no tiene por qué aparecer en primer término en la consulta original (Q). De esta forma se consigue que el conjunto de documentos resultantes se aproxime en gran medida a la idea de relevancia que posee el usuario.

- El segundo método se denomina *Pseudo Relevance Feedback (PRF)*, y su mayor ventaja es que no requiere la intervención del usuario, evitando así a éste la farragosa tarea de seleccionar de un conjunto de documentos aquellos que considera relevantes. Para ello se consideran como documentos relevantes a todos aquellos documentos que son devueltos en las primeras posiciones del ranking una vez resuelta la consulta original por el motor de búsqueda. Sin embargo el problema principal de esta aproximación, es que en algunos casos la nueva consulta Q' degrada la calidad de la respuesta respecto a la consulta original Q. Esta degradación de los resultados ocurre cuando el conjunto de documentos considerados relevantes, es decir aquellos devueltos en primer término por el SRI, no son suficientemente informativos o están claramente desviados del concepto que el usuario intentó expresar a través de su consulta original (Ruthven y Lalmas (2003)).

Es en relación a las limitaciones que muestran los distintos métodos de expansión de consultas en ausencia de información del usuario *(PRF)*, donde surge una de las principales motivaciones de este trabajo de tesis. Un estudio reciente (He y Ounis (2009)), que analiza en detalle el rendimiento de los distintos métodos de expansión de consultas mediante *PRF*, concluye que su rendimiento depende en gran medida de la calidad de la respuesta inicial del SRI (He y Ounis (2009)).

1.3. Predicción de la calidad de consultas

Una posible aproximación para resolver los problemas inherentes a la expansión de consultas sin intervención del usuario consiste en predecir la calidad de la respuesta inicial. Con esta información la expansión podría llevarse a cabo solo en el caso en el que la respuesta inicial devuelta por el SRI fuera de suficiente calidad. Es decir, teniendo la capacidad de detectar la calidad de la respuesta dada a una consulta, se podría realizar una aplicación de los métodos de expansión de consultas más adecuada.

La predicción de la calidad de la respuesta dada por un SRI (*Query Performance Prediction (QPP)*) ha sido estudiada durante los últimos años en el entorno de la recuperación de información (Carmel et al. (2006); Hauff (2010); Carmel y Yom-Tov (2010)). El objetivo principal de este tipo de métodos, es el de estimar la calidad de la respuesta de un SRI a una consulta especificada por un usuario, en ausencia de juicios de relevancia. Existen

diversas razones por las que un SRI, en algunos casos, es incapaz de devolver una respuesta adecuada para una consulta, incluyendo causas como la ambigüedad de la consulta, la falta de cobertura sobre el asunto que describe la consulta o simplemente la existencia de errores ortográficos en la consulta (Carmel et al. (2006)).

La aplicación de este tipo de tecnologías no se limita tan solo al caso de la expansión de consultas. Por ejemplo, el ámbito de los metabuscadores se vería beneficiado por este tipo de técnicas con la posibilidad de seleccionar la respuesta más adecuada de un conjunto de buscadores distintos.

Otra posible aplicación estudiada recientemente (Baeza-Yates et al. (2009)), se refiere a la posibilidad de que los grandes buscadores subdividan sus índices con el objetivo de ahorrar costes. Estos subíndices reflejarían un idioma o temática distinta y dependiendo de su tamaño el acceso a ellos incurriría en un coste distinto. De esta forma, con la capacidad de predecir la calidad de una respuesta, se podría establecer la política por defecto de usar aquellos subíndices que incurrieran en costes menores. En el caso de que la respuesta obtenida a partir de estos subíndices fuera de baja calidad, se realizaría la misma petición a índices de mayor calidad y que generalmente supondrían un coste asociado más elevado. Recientemente han aparecido trabajos que hacen uso de métodos de predicción pero enfocados a estimar la calidad dada por un sistema de recomendación (Bellogín (2011); Bellogín et al. (2011)).

Una de las primeras referencias a *QPP* aparece en el año 2002 en los trabajos desarrollados por Cronen-Townsend et al. (2002). En este trabajo se propone como estimador de la calidad de una respuesta el método de predicción *Clarity Score*, que mide la divergencia entre el modelo de lenguaje de los documentos recuperados con la consulta original y el modelo de lenguaje construido a partir del de búsqueda. Así, un alto grado de divergencia indicará que ambos vocabularios difieren en gran medida y que por tanto la consulta es lo suficientemente específica como para poder ser resuelta con ciertas garantías. Aunque *Clarity Score* mostró resultados muy prometedores en términos de predicción, tenía como principal desventaja su coste computacional que hacía difícil su aplicación en un entorno real.

La evaluación del rendimiento de un método de predicción para un grupo de consultas, se calcula usando distintas medidas de correlación entre los valores estimados por el método de predicción y la calidad de la respuesta dada a cada consulta, siendo está calidad evaluada de acuerdo con los juicios de relevancia.

La calidad de la respuesta que obtiene una consulta, se mide habitualmente haciendo uso de la denominada precisión media *Average Precision (AP)*, que es considerada una de las medidas estándar en el campo de la RI.

En la actualidad existen numerosos trabajos en el campo de la predicción del rendimiento de una consulta, pero en general aparecen dos principales limitaciones en estos métodos: en relación con el grado de acierto en las predicciones y al coste computacional necesario para llevar acabo

dichas predicciones. Así, por un lado aquellos métodos con un mayor poder de predicción suelen conllevar un elevado coste computacional, lo que puede dificultar su integración en un entorno real. Por otro lado los métodos con menores requisitos computacionales, no son capaces de mostrar un rendimiento suficientemente robusto en términos de predicción (Carmel y Yom-Tov (2010), pág. 16).

Por tanto se plantea la necesidad de desarrollar nuevos métodos de predicción que cubran las carencias anteriormente descritas. La propuesta a desarrollar en esta tesis se basa en algunas de las conclusiones extraídas del trabajo realizado por Araujo et al. (2010). Aunque este trabajo no estaba dedicado a la predicción de la calidad de consultas, en su desarrollo se observaron ciertos indicios de que algunos de los estadísticos propuestos, podrían ser aplicados al campo de la predicción de la calidad de una respuesta a modo de estimadores. Estos estimadores se calculan a partir de los distintos valores de relevancia, o pesos, asignados por las funciones de ranking a los documentos y eran utilizados como funciones de ajuste de un algoritmo genético, demostrando su utilidad en el campo de la expansión de consultas de carácter morfológico.

Por tanto, un primer paso para el desarrollo de una nueva familia de métodos de predicción, estaría basado en el desarrollo de un estudio detallado de la capacidad de predicción de estos estadísticos, que son calculados a partir de los pesos asignados a los documentos por la función de ranking.

Otra línea de investigación que se considera de gran importancia en el campo de la predicción de la calidad de las consultas, es aquella relacionada con la evaluación de la calidad de estos métodos. En el presente, dicha evaluación se lleva a cabo basándose en diferentes coeficientes de correlación. Sin embargo, no existe actualmente un consenso amplio sobre qué coeficiente es el más adecuado. Además, como se puede observar a partir de los resultados publicados por Hauff (2010) (cáp. 2 y cáp. 3), existe una gran similitud entre los valores de correlación obtenidos por los distintos métodos de predicción. Esto dificulta sobremanera una correcta comprensión del rendimiento real de dichos métodos de predicción y por tanto es difícil conocer su idoneidad para distintas aplicaciones.

Por tanto, existe una clara necesidad de realizar un estudio profundo de las características de la evaluación de métodos de predicción basada en la correlación, que permita concluir como de efectivas son estas medidas para describir en detalle la potencialidad de estos métodos frente a una tarea específica.

1.4. Objetivos

Los objetivos principales a desarrollar dentro de este trabajo de tesis, se centran en el campo de la predicción de la calidad de consultas. Dentro

de este campo, se plantearon inicialmente tres líneas principales de investigación: (a) analizar si los valores de relevancia asignados por una función de ranking podrían ser indicativos de la calidad de una consulta; (b) analizar las principales limitaciones que aparecen con el marco de evaluación basada en coeficientes de correlación con el que se evalúan los métodos de predicción; (c) estudiar la potencial mejora de rendimiento en el campo de la expansión de consultas con la introducción de técnicas de predicción.

Estos objetivos principales son los que dan estructura y sentido al conjunto de la tesis aquí presentada y se subdividen en las siguientes tareas de forma más específica:

- Realizar un estudio de las distintas técnicas de predicción de la dificultad de una consulta actuales, así como de los distintos enfoques existentes.

- Desarrollar una propuesta de técnica de predicción con el principal objetivo de superar algunas de las principales limitaciones de los métodos actuales, como mejorar la calidad de las estimaciones o disminuir el coste computacional asociado. La aproximación principal para el desarrollo de esta nueva técnica de predicción, se basa en el cálculo de la desviación estándar de los diferentes valores de relevancia asignados por la función de ranking al conjunto de documentos recuperados.

- Realizar un análisis crítico del marco de evaluación actual que se emplea para medir el rendimiento de los distintos métodos de predicción. El objetivo de esta tarea es comprobar si existe una relación clara entre los resultados obtenidos al evaluar métodos de predicción mediante correlación, y la utilidad de estos métodos para su posible aplicación en distintos campos.

- Definir medidas de evaluación alternativas, que enfaticen las diferencias en términos de rendimiento o robustez de las predicciones obtenidas por distintos métodos y que por tanto faciliten la selección justificada de un método de predicción para un entorno concreto o problema específico.

- Evaluar tanto los métodos de predicción propuestos como el marco de evaluación desarrollado para distintas colecciones y métodos de predicción. Esta tarea tiene como objetivo la evaluación tanto cuantitativa como cualitativa de la utilidad del nuevo marco de evaluación, propuesto para un subconjunto de métodos de predicción representativos.

- Realizar un análisis que permita establecer la utilidad de la aplicación de distintas técnicas de predicción dentro del campo de la expansión de consultas.

- Finalmente y en base a las conclusiones obtenidas con el trabajo realizado, se plantearan líneas de trabajo futuro consecuentes con los resultados obtenidos.

1.5. Metodología

La metodología básica que se utiliza en este trabajo de tesis es el denominado paradigma *Cranfield* (Voorhees (2002)), en el que se apoyan conferencias tan importantes como *TREC*, *CLEF* o *NTCIR*. Este paradigma ha dado lugar a la denominada evaluación de laboratorio dentro del campo de recuperación de información.

Este tipo de evaluación ha ido perfeccionándose a lo largo de los años, desde los primeros experimentos *Cranfield 2* desarrollados en la década de los sesenta, dando lugar a una metodología de evaluación que, aunque no exenta de críticas, se considera como un marco robusto y fiable.

El objetivo principal de este paradigma o marco de investigación, se centra en facilitar la comparación del rendimiento que muestran diferentes sistemas de recuperación de información a la hora de resolver distintas tareas. Esta metodología, en contraposición a una evaluación basada en usuarios, intenta reducir al mínimo la intervención de estos con el único fin de disminuir los costes y la complejidad asociada a una evaluación basada en un conjunto heterogéneo de usuarios.

Los componentes básicos que permiten realizar evaluaciones de distintos sistemas de forma sistemática y robusta son: el uso de colecciones estándar y suficientemente representativas del problema al que son aplicadas; un conjunto de consultas sobre esta colección; un conjunto de juicios de relevancia que permitan dada una consulta específica establecer, al menos, que documentos son relevantes para dicha consulta; y una o varias medidas de evaluación a través de las cuales definir el desempeño objetivo de los distintos sistemas de recuperación que son evaluados.

Las principales críticas que se realizan al paradigma *Cranfield*, son consecuencia de que especifica como ciertas un conjunto de supuestos con el objetivo de facilitar la comparación entre sistemas. El conjunto de consideraciones que se establecen a priori y que en muchos casos se alejan de la realidad se pueden resumir en los siguientes puntos:

- A priori todos los documentos dentro de la colección son equivalentes, se considera que todos son igualmente relevantes.

- El hecho de que un documento sea juzgado como relevante es independiente respecto a la relevancia de otros documentos distintos.

- La necesidad de información expresada por el usuario a través de una consulta es estática. Es decir, la necesidad de información no evolucio-

na según se va adquiriendo conocimiento con la observación de otros documentos de la colección.

- Se asume que el conjunto de juicios de relevancia realizado es representativo para el conjunto de usuarios posibles de los sistemas evaluados.

- Se considera que el conjunto de documentos considerados relevantes dada una consulta según los juicios de relevancia es completo, es decir no existen en la colección otros documentos que pudieran ser considerados como relevantes.

De forma específica para el conjunto de experimentos que se realizarán en esta tesis se considera el concepto de relevancia como binario, por lo tanto un documento será relevante o no pero sin tener en cuenta distintos grados de relevancia. Además se establece que cualquier sistema de información utilizado a lo largo de esta tesis, estará enfocado de manera específica a la denominada recuperación de información *"ad-hoc"*.

La recuperación *"ad-hoc"* está considerada como el tipo de recuperación más estándar dentro de la RI. En este tipo de recuperación el objetivo del sistema, es recuperar, a partir de una colección, el subconjunto de documentos más relevante respecto a una necesidad informativa suministrada por el usuario, considerando que la colección permanece estática, mientras que las consultas suministradas al sistema por el usuario variarán a lo largo del tiempo (Baeza-Yates y Ribeiro-Neto (1999), pág. 21,22).

El conjunto de colecciones que se utilizarán para evaluar la validez de los distintos enfoques propuestos engloban corpus clásicos que se utilizan frecuentemente en las tareas de *TREC* como *TREC Vol. 4 +5*, *WT10G* y *GOV2*, junto con los distintos conjuntos de consultas utilizadas en tareas como las de *Robust2004*, *Web Track* o *Terabyte Track*.

En relación a la evaluación de métodos de predicción, como se aplica habitualmente, se utilizarán los coeficientes de correlación *Pearson* y *Kendall*.

1.6. Estructura de la tesis

Los contenidos de esta tesis se organizan en torno a los siguientes capítulos:

- Capítulo 2: En este capítulo se realiza un recorrido por la literatura relacionada con la predicción de la calidad de consultas, describiendo el marco de investigación dentro de este campo así como el conjunto de métodos más destacados que han aparecido hasta ahora. Además se enumera y describe en detalle las diferentes herramientas utilizadas en este trabajo de tesis, con el objetivo de facilitar la comprensión de las distintas técnicas y aportaciones principales de esta investigación.

- Capítulo 3: El capítulo 3 está dedicado al estudio y análisis de la propuesta de predicción de la calidad de una consulta en base a los valores de relevancia asignados por una función de ranking. La técnica de predicción propuesta y una evaluación en detalle del rendimiento que alcanza dicha aproximación conforman el contenido principal que aparece en este capítulo.

- Capítulo 4: En este apartado de la tesis aparece un análisis crítico del marco de evaluación aplicado actualmente al campo de la predicción de la calidad de consultas, destacando aquellas limitaciones que aparecen en dicho marco. Además se introduce un novedoso marco de evaluación con el principal objetivo de superar algunas de las limitaciones encontradas, consecuencia de una evaluación basada únicamente en coeficientes de correlación.

- Capítulo 5: El capítulo 5 se enfoca a la aplicación del uso de métodos de predicción en un campo como el de la expansión de consultas en ausencia de información de usuario. De esta forma se aplica el método de predicción introducido en el Capítulo 3 y se estudia la influencia en términos de rendimiento que produce su uso en un entorno de expansión selectiva frente al más clásico de expansión sistemática, donde se expanden el conjunto de todas las consultas.

- Capítulo 6: Finalmente, el último capítulo se centra en destacar las principales conclusiones obtenidas como consecuencia de la realización de este trabajo, junto a los posibles desafíos futuros que se plantean en el campo de la predicción de la calidad de las consultas.

Each man must reach his own verdict, by weighing all the relevant evidence.

Leonard Peikoff.

Capítulo 2

Antecedentes

El objetivo de este capítulo es introducir las distintas técnicas que intervienen en el trabajo a desarrollar, de forma que se facilite una correcta comprensión del conjunto de tareas que se pretende abarcar en esta tesis. Se realizará una breve introducción a algunos de los modelos de recuperación de información y a sus componentes fundamentales, así como una descripción del marco de evaluación típico de los sistemas de recuperación de información. Posteriormente y ya dentro del campo del refinamiento de las consultas, se realizará un análisis del trabajo de Araujo et al. (2010), algunos de cuyos resultados han intervenido en la motivación de esta tesis. Finalmente nos centraremos en el campo de la predicción de la dificultad de consultas que constituye el núcleo principal de esta tesis.

2.1. Modelos de recuperación de información

Esta sección tiene como objetivo describir algunos de los distintos modelos que se aplican en la actualidad en recuperación de información *"ad-hoc"*, esto es, recuperar un conjunto de documentos relacionados con una necesidad informativa. Dicha necesidad informativa expresa un asunto específico del cual el usuario desea recibir información. Es importante establecer una clara distinción entre el concepto de necesidad informativa respecto al de consulta. Mientras que el primero describe el asunto en el cuál el usuario muestra interés, la consulta o *"query"* es la entrada que se suministrará al sistema, para cubrir dicha necesidad informativa. Por tanto una necesidad informativa puede venir representada por distintas consultas y ser respondida o cubierta por diferentes documentos.

El concepto clave que desarrollan los diversos modelos que se describen a continuación, es la modelización de una función ideal que cuantifique cómo de relevante es un documento para una consulta expresada en forma de un conjunto de términos. Cualquier aproximación que intente resolver este problema se enfrenta a priori a una dificultad básica: el hecho de que

la relevancia es un concepto totalmente subjetivo y por tanto dependiente del usuario. Devolver al usuario un documento específico que contenga la información que él pretende conseguir es una tarea que no está resuelta, ya que aún no es posible entender cómo funciona el proceso a través del cual se expresa una necesidad de información.

Antes de iniciar un recorrido por los distintos modelos propuestos, es necesario describir ciertos pilares sobre los que estos modelos se sustentan.

A la hora de calcular la relevancia de un documento, un primer paso necesario es definir cómo se representa dicho documento dentro del modelo aplicado. En recuperación de información la representación más extendida para un documento es la basada en *"bag of words"*, o bolsa de palabras, según la cuál se representan los documentos en base al conjunto de unidades básicas que conforman cada documento.

Las unidades básicas corresponden con las palabras que contienen los documentos a las cuales se aplican previamente una serie de procesos. Estos procesos suelen consistir en la eliminación de las denominadas palabras vacías, eliminación de signos de puntuación, transformación a minúsculas, lematización u otras. Una vez realizados estos procesos se obtienen las unidades básicas o términos que en su conjunto conformarán la representación del documento. Este tipo de representación ignora características propias del documento como el orden de las unidades básicas, o la gramática específica del documento.

Otra cuestión importante a la hora de introducir los distintos modelos aplicados a la recuperación de información es cómo se puede medir la efectividad de un sistema de este tipo, es decir la calidad de los resultados devueltos. En general en RI se definen dos estadísticos básicos que serán los que indicarán el rendimiento de un sistema para unas necesidades específicas. Estos dos estadísticos son la precisión y la cobertura:

- **Precisión**, cuyo objetivo es medir qué fracción de los documentos devueltos son realmente relevantes dada una necesidad informativa.

- **Cobertura**, cuyo objetivo es medir cuántos de los documentos relevantes presentes en la colección han sido recuperados por nuestro sistema.

Estas medidas son la base fundamental para evaluar cualquier sistema de recuperación, y partiendo de ellas se han desarrollado nuevas funciones de evaluación más sofisticadas que se tratarán en detalle posteriormente en la Sección 2.2.1.

Una vez descritos los conceptos anteriores, comunes a las distintas aproximaciones, ya es posible abordar la tarea de describir los distintos modelos de recuperación de información.

2.1.1. Modelo de espacio vectorial *(VSM)*

El modelo de espacio vectorial, o *VSM* por sus siglas en inglés, parte de la idea de considerar a las consultas y a los documentos como entidades equivalentes; esto implica que las consultas, al igual que los documentos, se modelizarán según el esquema de bolsa de palabras. Este tipo de consultas en las que los términos no se encuentran conectados mediante operadores booleanos, se denominan de *"texto libre"* o *"free-text queries"* y son el tipo de consultas usado con mayor frecuencia en los buscadores generalistas que podemos encontrar habitualmente en la Web.

Este modelo ha sido uno de los más aplicados en el campo de recuperación de la información durante los últimos años y fue descrito por primera vez por Salton et al. (1975), donde se describía una representación vectorial de los documentos y las consultas. A partir de esta representación vectorial, Salton define una medida de similitud entre el vector que representa la consulta y el vector de un documento. La intuición en la que se sustenta este modelo es que los distintos temas que abarca un documento vienen determinados por las palabras que contiene. Si se representa un documento como un vector cuyas componentes representan la importancia de cada palabra en el documento y se realiza el mismo proceso con la consulta, se podría aplicar una medida de similitud que indicara la cercanía del "contenido" de ambos. A partir de este valor de similitud se puede construir fácilmente una lista ordenada de documentos en base al grado de cercanía de cada uno de los documentos en relación a otro documento base o consulta.

De un modo aproximado se puede describir el modelo de espacio vectorial, como un método que calcula la relevancia parcial de cada término en la consulta respecto al documento. A continuación se suman todos los valores correspondientes a los términos de la consulta, de tal forma que se obtiene un número que representa el valor de relevancia de dicho documento para la consulta especificada. Una de las claves de la correcta ordenación de los resultados vendrá dada por las distintas funciones que se utilicen para calcular la importancia de un término en un documento. Estas funciones se basan principalmente en estadísticos que se calculan a nivel de documento y de colección. Los estadísticos se suelen denominan en la literatura como "esquemas de pesado". Como se comprobará posteriormente (Sección 2.4), algunos de los componentes de los denominados "esquemas de pesado" son utilizados como métodos para predecir la calidad de consultas.

Representación

Con la representación del modelo vectorial todo documento de la colección se describe en base a un espacio vectorial de dimensión $n \times m$, siendo n el número de documentos en la colección y m el número de términos distintos que aparecen a lo largo de toda la colección. Así, cada documento

vendrá representado por un vector de m componentes, donde cada componente tendrá un valor numérico que indica la importancia del término correspondiente en el documento. Por ejemplo, asignando uno en el caso de que el término aparezca en el documento y cero en caso contrario:

	$\mathbf{t_1}$	$\mathbf{t_2}$	...	$\mathbf{t_m}$
$\mathbf{d_1}$	1	0	...	1
$\mathbf{d_2}$	0	0	...	1
...	...	...	...	...
$\mathbf{d_n}$	1	1	...	1

A su vez una consulta q se representa como un vector en el mismo espacio multidimensional utilizado para representar los documentos de la colección:

	$\mathbf{t_1}$	$\mathbf{t_2}$	...	$\mathbf{t_m}$
$\mathbf{q}$	0	0	...	1

Una vez definida la representación que se usará, se debe definir a su vez una medida de similitud entre ambos vectores en el espacio multidimensional descrito.

Medida de similitud

La aproximación más utilizada como medida de similitud es el coseno del ángulo que forman ambos vectores. Este coseno se calcula haciendo uso del producto escalar entre ambos vectores, de tal forma que si tenemos el vector de la consulta $\vec{V}(q)$ y el vector del documento $\vec{V}(d)$, la similitud entre ambos vectores se define como:

$$sim(q,d) = \frac{\vec{V}(q) \cdot \vec{V}(d)}{|\vec{V}(q)||\vec{V}(d)|}$$

donde el numerador representa el producto escalar entre los dos vectores y el denominador es el producto de los módulos de ambos vectores.

Si denotamos las componentes del vector que representa a un documento como w_i y las componentes del vector de la consulta como q_i, el valor de relevancia de un documento d respecto a una consulta q, se define como:

$$sim(q,d) = \frac{\sum_{i=1}^{m} q_i w_i}{\sqrt{\sum_{i=1}^{m} q_i^2}\sqrt{\sum_{i=1}^{m} w_i^2}} \tag{2.1}$$

La medida de similitud tal y como se ha definido de forma teórica en la ecuación 2.1, presenta un grave problema como consecuencia del gran número de dimensiones de los vectores y por tanto del coste computacional necesario para el cálculo del coseno. En la actualidad es común utilizar colecciones que contengan millones de documentos con miles, o incluso cientos

de miles, de términos, razón por la cual se suelen utilizar distintas aproximaciones que faciliten los cálculos de similitud.

La simplificación más típica de la ecuación 2.1, se obtiene utilizando como factor de normalización el número de términos que contenga el documento que se está evaluando en lugar del producto de los módulos. Así se elimina la necesidad de calcular el módulo del vector que corresponde al documento, cuyo cálculo implica un elevado coste computacional.

$$sim(q,d) = \frac{\sum_{i=1}^{m} q_i w_i}{\sqrt{numTerm}}$$

siendo $numTerm$ el número de términos presentes en el documento d. Este método permite un cálculo de la similitud considerablemente más rápido respecto a la ecuación original, aunque tiende a favorecer ligeramente, es decir, situar más arriba en el ranking, a aquellos documentos más largos en comparación con la aplicación del coseno clásico. Esta aproximación aparece con bastante frecuencia en aquellos sistemas que implementan el modelo de espacio vectorial como función de ranking.

La frecuencia como factor de relevancia

Una vez descrita la función de similitud, se debe definir una ecuación que calcule cómo de relevante es un término para un documento. El valor obtenido a partir del esquema seleccionado será el que se utilice en la ecuación 2.1 como w_i. El concepto de esquema de pesado surge a partir del sistema *SMART* (Salton y Lesk (1965)) y permite realizar, de forma sencilla, cambios en la función de ranking de un sistema de recuperación mediante el ajuste de la ecuación a partir de la cuál se obtiene el valor w_i.

Una primera aproximación se basa en calcular este valor a partir de la frecuencia del término en el documento. Esto es, cuantas más veces aparezca un término en un documento más representativo será para dicho documento. Esta primera aproximación se denomina frecuencia del término t en el documento d y se denota como $tf_{t,d}$.

Una primera consecuencia de la utilización de este esquema es el hecho de que los términos más comunes en un idioma gozarían de gran importancia en el cálculo de relevancia. Sin embargo, es un hecho conocido que los términos de mayor frecuencia en un idioma suelen ser palabras con un contenido semántico mínimo, sirvan como ejemplo los artículos o preposiciones en castellano. Este tipo de palabras se suelen denominar en recuperación de información palabras vacías, y como consecuencia de su alta frecuencia suelen ser eliminadas de los índices, de forma que no desvirtúen los estadísticos que se puedan obtener a partir de los documentos.

Otra restricción que se suele aplicar a la frecuencia de un término en un documento, se refiere a evitar el cálculo de la relevancia de un término en un documento directamente en base a su frecuencia de aparición. Sin

esta restricción se consideraría que la importancia de un término crece de forma lineal con la frecuencia. Esto es, un término con frecuencia $n + 1$ es más importante que un término con frecuencia n, siendo este incremento de la relevancia equivalente para cualquier $n \in \mathbb{N}$. Esta concepción es intuitivamente errónea ya que no parece lógico que el incremento de relevancia sea equivalente al pasar de una frecuencia de un término de dos a tres, que al pasar de una frecuencia quince a dieciséis. Controlar la influencia de la frecuencia en relación a la relevancia se denomina saturación de la frecuencia, y consiste en eliminar el comportamiento lineal entre relevancia de un término y su frecuencia. Existen diversas aproximaciones para resolver este problema pero dos ejemplos muy extendidos consisten en la aplicación de la raíz cuadrada o del logaritmo a dicha frecuencia.

Otro esquema de pesado bastante común basado en la frecuencia consiste en normalizar todas las frecuencias en un documento con la máxima frecuencia encontrada en dicho documento. Así, w_i vendría dado por la siguiente ecuación:

$$w_i = \frac{tf_{t,d}}{tf_{max(t,d)}}$$

El objetivo principal de este esquema es mitigar el efecto que tiene sobre la frecuencia la longitud de los documentos. Se puede intuir que los documentos más largos tienden a tener frecuencias más altas. Esto no solo ocurre por el hecho de contener más términos, lo que hace más probable que aparezcan términos repetidos, sino por la idea de que los documentos más largos tienden a describir el mismo contenido con mayor verbosidad.

Estadísticos de colección como factores de relevancia

Los esquemas descritos hasta ahora son esquemas basados en la frecuencia de términos, es decir en lo que se denomina estadísticos del documento. El problema de cuantificar la relevancia de un término solo en base a estadísticos del documento se debe a considerar que todos los términos de la consulta tienen la misma importancia o grado de especificidad respecto a la colección. En realidad, algunos términos de la consulta tendrán poca o ninguna relevancia. Por ejemplo, en una colección que trate sobre un dominio concreto habrá términos que aparecerán mucho a lo largo de la colección, pero que por contra no serán especialmente discriminantes. Se puede imaginar la palabra "medicamento" o "cura" en una colección de tratados médicos. Aún así la aparición de estos términos en la consulta, provocará que los documentos que los contengan adquieran un valor de relevancia muy elevado por tener una frecuencia muy alta, aunque no contengan otros términos de la consulta mucho más discriminantes. La utilización de estadísticos de la colección para realizar el cálculo de relevancia de un término, fue desarrollado principalmente por Sparck Jones (1972), y ha sido uno de los grandes hitos

en el campo de la recuperación de información debido al salto de calidad que se observa en los resultados con su uso.

El hecho de introducir en el esquema de pesado estadísticos relativos a la colección, nos permitirá atenuar en gran medida el efecto producido por los términos muy comunes. Este objetivo se puede conseguir mediante la introducción de un nuevo factor que disminuya la importancia de un término de la consulta cuanto más frecuente sea éste a lo largo de la colección.

Existen dos estadísticos principales que se usan para este objetivo:

- Frecuencia en la colección, es decir, el número de veces que aparece un término t a lo largo de la colección y se denota como cf_t.

- Frecuencia por documento, que calcula el número de documentos de la colección en los que aparece el término t, y se denota como df_t.

Aunque pueden parecer medidas similares, se ha observado que en algunos casos el uso de df_t, suministra información más adecuada al esquema de pesado que cf_t. Esto ocurre simplemente por el hecho de que el estadístico obtenido a partir del valor de cf_t, puede estar desvirtuado en el caso de que exista en la colección uno o varios documentos con una frecuencia del término t muy elevada. En cambio los valores de df_t suelen describir de forma más fidedigna la representatividad de t en la colección (Manning et al. (2008), pág. 118).

El valor de df_t no se aplica directamente sobre la función de pesado, sino que a partir de éste se calcula la denominada frecuencia inversa de documento idf_t. El valor de idf_t se puede expresar de diversas formas pero la forma básica de dicha expresión se suele encontrar como aparece en la siguiente ecuación:

$$idf_t = log\frac{N}{df_t} \tag{2.2}$$

donde N es el número de documentos de la colección.

Se puede demostrar de forma teórica que el cálculo de idf_t modela la probabilidad de que un término t aparezca en un documento relevante para dicho término. El desarrollo matemático de este hecho, así como del efecto de la aplicación de idf_t a un esquema de pesado se puede encontrar en Robertson (2004).

Con la utilización de los estadísticos de la colección, el esquema de pesado de un término t, quedará definido en base a la combinación del valor obtenido a partir de la frecuencia de t, junto a su frecuencia inversa de documento o idf_t. Esta combinación se suele denominar esquema de pesado *TF-IDF*.

La combinación de ambas medidas nos permitirá obtener valores de relevancia para un término, que de forma intuitiva serán:

- Altos, cuando el término t tenga una frecuencia alta en un número pequeño de documentos.
- Medios, cuando t tenga una frecuencia baja en un documento, o tenga una frecuencia alta a lo largo de la colección.
- Bajos, siempre que el término sea muy común a lo largo de la colección.

Una vez desarrollados algunos de los esquemas de pesado existentes para la ecuación 2.1 basados en la combinación *TF-IDF*, esta ecuación se podría redefinir en función de un esquema de pesado específico para w_i, obteniéndose, por ejemplo, una función de similitud como la siguiente:

$$sim(q,d) = \sum_{t\ en\ c} \frac{q_t w_t}{\sqrt{numTerm}}$$

donde $numTerm$ es el número de términos de d, y

$$w_t = tf_{t,d} \cdot idf_t$$

obteniéndose el valor de tf_t e idf_t a partir de

$$tf_{t,d} = 1 + log(frec_{t,d}),$$
$$idf_t = log\frac{N}{df_t},$$

siendo esta expresión final una de las clásicas que se aplican en diversos sistemas de recuperación de información que hacen uso del modelo de espacio vectorial.

2.1.2. Modelo probabilístico

El modelo probabilístico trata de modelizar la incertidumbre existente en el hecho de si un documento será relevante o no, y con qué grado, respecto a una necesidad informativa. La aplicación de los fundamentos de la teoría de la probabilidad se adaptan perfectamente a problemas en los que se debe modelar incertidumbre, como es el caso de la probabilidad de que un documento sea relevante para una consulta dada. En Fuhr (1992) o Crestani et al. (1998), se puede encontrar una descripción teórica muy amplia de las distintas aproximaciones al modelo probabilístico que se aplican en recuperación de la información. En este trabajo se describirá en mayor detalle la aproximación a través de la que se obtiene la función de ranking *BM25* que se utilizará frecuentemente en esta tesis.

Al igual que en los modelos descritos con anterioridad, el modelo probabilístico utiliza los conceptos de necesidad informativa y documentos. Partiendo de representaciones para ambos, el objetivo de un sistema será el de

determinar como de bien cubren los documentos las necesidades informativas propuestas por los usuarios. Para resolver este problema, el sistema deberá ser capaz de estimar en cierta medida, si un documento posee información relevante para una consulta dada. De esta forma el objetivo de la utilización de medidas probabilísticas es tratar de calcular el grado de similitud entre una consulta y un documento basándose en la probabilidad de que un documento d sea relevante para una consulta q.

Principio de ranking probabilístico (PRP)

Si partimos de un documento d y una consulta q, se puede definir la relevancia R como una variable aleatoria que tomará el valor uno en caso de que d sea relevante para dicha consulta y cero en caso contrario. Utilizando el modelo probabilístico, los documentos serán ordenados según su relevancia que vendrá estimada por la probabilidad de que la variable aleatoria R tome el valor uno, dados q y d, es decir una probabilidad condicionada que se denota como $P(R|d,q)$.

El Principio de Ranking Probabilístico o PRP fue descrito y justificado por van Rijsbergen (van Rijsbergen (1979), pág. 113-114). Este principio propone un marco teórico que es la base del modelo probabilístico. La idea básica que subyace del PRP es que para recuperar un conjunto de documentos de forma óptima, estos deben ser ordenados de acuerdo a su probabilidad de ser relevantes. Es decir, un documento d_m deberá ser recuperado de la colección si y solo si, no existe en la colección otro documento d_j con una probabilidad mayor de ser relevante:

$$P(R|d_m,q) \geq P(R|d_j,q)$$

de tal forma que es posible obtener un conjunto ordenado de documentos en base a su relevancia.

Best Match 25 (BM25)

La función de ranking *BM25* corresponde a la familia de funciones de ranking probabilísticas cuyos fundamentos se apoyan en la modelización descrita en el denominado modelo de independencia binario *(BIM)*, que aparece en Robertson y Sparck Jones (1988).

El objetivo del modelo de independencia binario, que utiliza como base las ideas principales del PRP, es el de conseguir un conjunto de documentos ordenados por relevancia, coincidiendo esta relevancia con la estimación de que un documento sea relevante para una consulta dada, esto es $P(R|d,q)$. Para estimar la probabilidad condicionada $P(R|d,q)$, el *BIM* plantea como hipótesis el hecho de que los términos se distribuyen de manera diferente entre los documentos relevantes y los no relevantes.

Si se representan los términos que aparecen en la colección de documentos C como un conjunto $T = \{t_1, t_2, \ldots, t_n\}$, entonces se pueden representar los términos que aparecen en un documento d como un vector binario $\vec{x} = \{x_1, x_2, \ldots, x_n\}$ siendo $x_i = 1$ si $t_i \in d$ y $x_i = 0$ en otro caso. Como consecuencia del desarrollo propuesto en Robertson y Sparck Jones (1988), se define el valor de relevancia de un término t en el documento d como:

$$RSV_d = \log \frac{N - n + 0{,}5}{n + 0{,}5} \tag{2.3}$$

siendo N el número total de documentos de la colección y n el número de documentos que contienen el término t, es decir df_t. Se debe destacar la similitud de la ecuación 2.3 con aquella que calcula el valor de IDF para un término en la ecuación 2.2.

Como evolución al *BIM* surgieron nuevos modelos que pretendían incluir en la función de similitud, características propias del documento que indicaran su relevancia. Así, estos modelos permiten establecer un esquema de ponderación probabilístico, que tiene en cuenta las frecuencias de los distintos términos o la longitud de los documentos, de forma similar a como se propone en los distintos esquemas de ponderación del modelo de espacio vectorial descritos previamente. Un ejemplo sería el modelo probabilístico no binario basado en los trabajos de Robertson que se describen en Robertson y Walker (1994).

Estos trabajos dieron lugar al desarrollo de la familia de los denominados *"best match"*, donde el máximo exponente en la actualidad es la función de ranking *BM25*, también conocida como *Okapi* en referencia al nombre del sistema donde fue implementada por primera vez. Este modelo en la actualidad, es uno de los de mayor éxito en el campo de la recuperación de información por la calidad de los resultados que se obtienen en comparación a otros modelos clásicos como por ejemplo el *VSM*.

El modelo parte de la premisa de incluir las frecuencias de los términos, más allá de la sola presencia o ausencia de estos como en el caso del *BIM*. La frecuencia de los términos en un documento debería ser un indicador de si el documento trata sobre el concepto representado por el término o no. Robertson plantea la hipótesis de que aquellos documentos que tratan sobre el concepto representado por un término, pertenecen a la elite de dicho término, y que el hecho de que un documento pertenezca o no a la elite tiene cierta correlación con la frecuencia de dicho término en el documento. En su aproximación se propone utilizar una doble distribución Poisson, para documentos relevantes y no relevantes, que modele el concepto de elite. Se pueden encontrar los detalles del desarrollo matemático en Robertson y Walker (1994).

Con la modelización de la elite de un documento mediante las distribuciones de probabilidad adecuadas, se obtiene una ecuación a partir de la

cual se puede calcular el valor de relevancia de un término en un documento. Esta función tendría las siguientes características:

1. Toma valor cero para frecuencias iguales a cero.
2. Es creciente monótona respecto al valor de la frecuencia del término en el documento.
3. Alcanza un máximo asintótico.
4. El máximo asintótico se acerca al valor que se obtiene a partir de la ecuación 2.3.

El siguiente paso consiste en desarrollar una función que se ajuste a las características anteriormente planteadas y que de forma simultánea refleje el efecto de la frecuencia en relación a la relevancia. Puesto que la función a desarrollar se acerca al valor de la ecuación 2.3, se puede definir en base a ésta. Si se denota w_1 al valor obtenido a partir de la ecuación 2.3, podemos definir el valor de relevancia como el producto entre w_1 y w, siendo w una nueva componente que dependerá del valor de tf. Robertson aproxima w a partir de la siguiente ecuación

$$w = \frac{tf}{k_1 + tf}$$

donde k_1 es un parámetro libre, que generalmente toma el valor 1,2 (He y Ounis (2007b)).

Se puede demostrar que el producto $w \cdot w_1$ toma valor cero cuando $tf = 0$, es creciente monótona respecto a tf y que alcanza un máximo asintótico. El valor donde se alcance dicho máximo asintótico dependerá del parámetro k_1. De esta forma, la utilización de este parámetro permite ajustar específicamente para una colección, el grado de saturación deseado de la función de ponderación que representa la frecuencia. Por ejemplo el uso de $k_1 = 1$ sería equivalente a obviar la frecuencia, esto es usar una ponderación binaria. Este tipo de aproximación a la saturación de la frecuencia mediante un parámetro, aparece junto a otras más clásicas, como la basada en el logaritmo o la raíz cuadrada, en la la Figura 2.1, donde k_1 toma el valor 1,2. La función resultante de multiplicar w y w_1 es de la forma:

$$RSV_d = \frac{tf}{k_1 + tf} \cdot \log \frac{N - n + 0{,}5}{n + 0{,}5}$$

Finalmente la función de ranking *BM25* incluye un factor que intenta representar el efecto que produce la longitud de un documento en su relevancia. Dicho factor se calcula basándose en una función que se denomina B. El objetivo de B es escalar la importancia de la frecuencia de un término

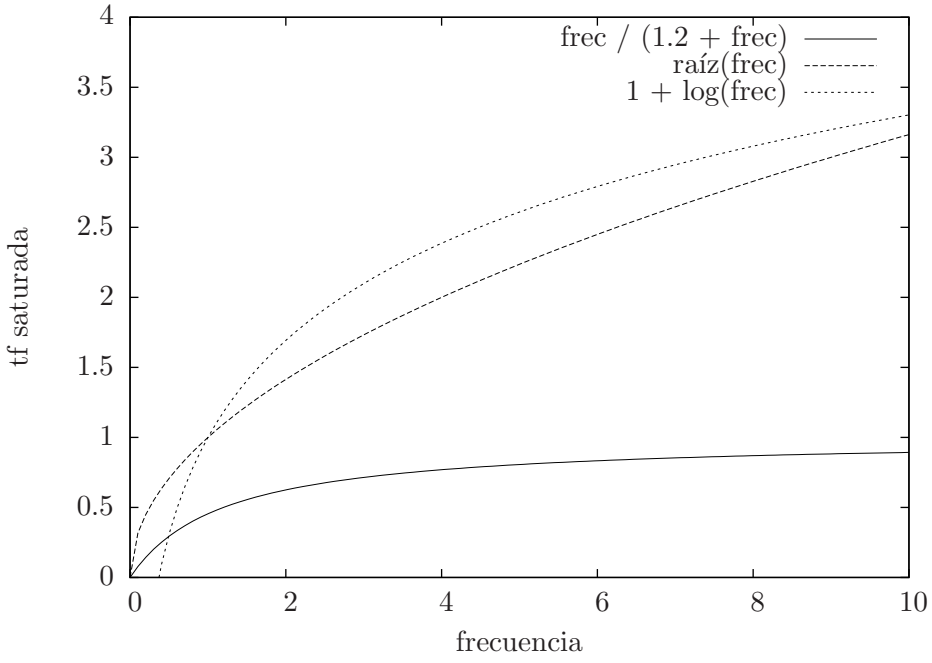

Figura 2.1: Comparación entre distintas aproximaciones de saturación de la frecuencia.

en el documento, en base a la longitud de éste. Así, la frecuencia de un término en el documento, tf, se divide por el resultado obtenido con la función B. Para el desarrollo de esta normalización se opta por la inclusión de un parámetro libre b, que deberá ajustarse específicamente para cada colección. Este parámetro tomará valores entre cero y uno, correspondiendo $b = 0$ a no tener en cuenta la longitud del documento y $b = 1$ a escalar el valor de la frecuencia con el total de la longitud del documento. En la siguiente ecuación se puede observar la formalización de la función B:

$$B = 1 - b + \left(b \cdot \frac{|D|}{avg} \right)$$

siendo $|D|$ la longitud del documento y avg la longitud promedio de los documentos pertenecientes a la colección. Se tiene entonces:

$$RSV_d = \frac{tf}{k_1 \left(1 - b + \left(\frac{|D|}{avg}\right)\right) + tf} \cdot \log \frac{N - n + 0{,}5}{n + 0{,}5} \tag{2.4}$$

En la actualidad *BM25* es una de las funciones de ranking con la que mejor rendimiento se obtiene dentro de aquellas que se fundamentan en el

modelo probabilístico, convirtiéndose en casi un estándar de facto en competiciones como *TREC*. El gran éxito de *BM25* viene provocado, en gran medida, por su gran flexibilidad, ya que es capaz de describir específicamente para cada colección, como afecta la frecuencia de un término y la longitud de los documentos al valor de relevancia correspondiente a un término y un documento. Esta flexibilidad se consigue con la optimización de sus parámetros de forma específica para la colección en la que se desee utilizar.

2.1.3. Modelos de lenguaje

Durante los últimos años las funciones de ranking basadas en modelos de lenguaje han gozado de un gran interés dentro de la comunidad de recuperación de información.

Los modelos de lenguaje son simplemente distribuciones de probabilidad sobre un conjunto de palabras o términos, de forma que sea posible calcular la probabilidad de aparición de una palabra o secuencia de éstas. En general, la mayoría de las aproximaciones, a modo de simplificación, asumen la independencia entre las apariciones de términos, es decir, estiman los modelos de lenguaje utilizando unigramas. Por tanto, se considera que la aparición de una secuencia de palabras es equivalente a la aparición de cada palabra de forma individual. Existen dos razones principales por las que se considera adecuada dicha simplificación: diversos experimentos a lo largo de los años han demostrado un rendimiento similar al obtenido con aquellas aproximaciones más complejas que utilizan n-gramas, en términos de la calidad de la respuesta (Zhai (2008), pág. 105); y permite además que las funciones de ranking que aplican dicha simplificación puedan ser calculadas en un tiempo razonable (Zhai (2008), pág. 7). De esta forma si V es el conjunto de palabras en un vocabulario y $w_1, w_2, ..., w_n$ es una secuencia de palabras cuya probabilidad de aparición se quiere calcular, donde $w_1, w_2, ..., w_n \in V$, dicha probabilidad vendría dada por la siguiente expresión:

$$p(w_1, ..., w_n) = \prod_{i=1}^{n} p(w_i)$$

Query Likelihood

La primera aproximación al uso de modelos de lenguaje para construir funciones de ranking fue propuesta en 1998 en el trabajo desarrollado por Ponte y Croft (1998). Este primer modelo denominado *Query Likelihood* era tan efectivo como aproximaciones más tradicionales como *VSM* o aquellas basadas en esquemas de pesado *TF-IDF*. La aproximación en la que se fundamenta esta nueva función de ranking, es ordenar los documentos según la probabilidad de que la consulta introducida por el usuario, hubiera sido generada a partir de los modelos de lenguaje correspondientes a cada uno

de los documentos de la colección. Esto es, ordenar los documentos según la probabilidad de que una consulta Q sea extraída de un modelo de lenguaje de un documento D. Así, dado el modelo de lenguaje θ_D construido a partir del documento D, los documentos se ordenaran en base al valor obtenido para $p(Q|\theta_D)$.

Una decisión importante a la hora de estimar el modelo de lenguaje para un documento, es definir qué tipo de modelo probabilístico es el más adecuado. Aunque en el trabajo original de Ponte y Croft se utilizaba una distribución de Bernoulli múltiple, la aproximación más exitosa es la que utiliza una distribución multinomial. La principal diferencia entre ambas aproximaciones es que la aproximación multinomial tiene en cuenta la frecuencia de aparición de los términos y no solo su presencia, como en el caso de la Bernoulli múltiple. Así, si asumimos un modelo de lenguaje multinomial y usamos un documento D para su estimación, la expresión resultante[1] que nos proporcionará la probabilidad de aparición de una palabra w vendrá dada por la siguiente ecuación:

$$P_{ml}(w|\theta_D) = \frac{frec_{w,D}}{|D|}$$

donde $frec_{w,D}$ es el número de veces que aparece la palabra w en el documento D, y $|D|$ es la longitud, o número de palabras, que aparecen en dicho documento. Utilizando la expresión anterior podemos formular la probabilidad de que la consulta Q sea generada por el modelo de lenguaje de D. Así dada una consulta Q que contiene los términos $q_1, q_2, ..., q_n$, esta probabilidad se estimaría según la siguiente expresión:

$$p(Q|\theta_D) = \prod_{i=1}^{n} p(q_i|\theta_D) \tag{2.5}$$

Suavizado

El problema clásico que ocurre con la expresión anterior es que al ser un producto de probabilidades parciales, en el caso de que uno de los términos de la consulta no aparezca en el documento, es decir su probabilidad estimada sea cero, el total de la expresión sería cero. Para solucionar este problema se han propuesto diferentes técnicas de suavizado, que incorporan estadísticos externos al documento de forma que la probabilidad de un término que no aparece en un documento sea mayor que cero. Aunque existe un gran número de técnicas para la realización de este suavizado, de forma experimental se ha observado que dos son las que ofrecen un mejor resultado sin suponer un coste computacional muy elevado (Zhai (2008), pág. 36). Ambos métodos se basan en la realización de una interpolación entre el modelo de lenguaje de la colección y el modelo de lenguaje del documento (Zhai y Lafferty (2001b)):

[1] Usando el estimador de máxima verosimilitud.

- *Jelinek-Mercer*: que realiza una interpolación usando un coeficiente λ para controlar la cantidad de suavizado, quedando la expresión resultante como:

$$p_\lambda(Q|\hat{\theta}_D) = \prod_{i=1}^{n} \left((1-\lambda)\frac{frec_{q_i,D}}{|D|} + \lambda \cdot p(q_i|C) \right)$$

  siendo $p(w|C)$ la probabilidad de que el término w aparezca en la colección C. Se puede observar que en el caso de que λ sea igual a cero se elimina el suavizado obteniendo una expresión equivalente a la ecuación 2.5, mientras que con λ igual a 1, la estimación se realiza utilizando como muestra solamente la colección.

- *Dirichlet*: como con el método de *Jelinek-Mercer* se aplica también un parámetro que describe el grado de interpolación, pero en este caso el valor de dicho parámetro y por tanto su efecto se define de forma relativa a la longitud del documento que se utiliza como muestra. Así la expresión queda como:

$$p_\lambda(Q|\hat{\theta}_D) = \prod_{i=1}^{n} \left(\frac{|D|}{|D|+\mu}\frac{frec_{q_i,D}}{|D|} + \frac{\mu}{\mu+|D|}p(q_i|C) \right)$$

  Si comparamos esta expresión con el método de *Jelinek-Mercer*, podemos observar que ambos métodos serían equivalentes si fijamos $\lambda = \frac{\mu}{\mu+|D|}$. Por lo tanto la aproximación de *Dirichlet* especifica un menor suavizado para aquellos documentos de mayor longitud, lo cual se ajusta a la intuición de que aquellos documentos más largos deberían ser mejores estimadores. Así en el caso extremo de que $|D| \to \infty$, se eliminaría de la interpolación la información de la colección.

Divergencia *Kullback-Leibler (KL)*

Una aproximación distinta al uso de modelos de lenguaje para construir funciones de ranking, es la basada en la divergencia *Kullback-Leibler (KL)*. La divergencia *Kullback-Leibler* (Kullback y Leibler (1951)) calcula la divergencia entre dos distribuciones de probabilidad discretas según la siguiente expresión:

$$D_{KL}(P||Q) = \sum_{i}^{n} P(i) log \left(\frac{P(i)}{Q(i)} \right)$$

siendo n el número de distintos valores que pueden tomar las variables aleatorias P y Q.

Este modelo se fundamenta en los trabajos presentados por Lafferty y Zhai (Lafferty y Zhai (2001); Zhai y Lafferty (2001a)). En el modelo basado

en la divergencia KL se parte de dos modelos de lenguaje, uno para la consulta θ_Q y otro para cada documento θ_D. De forma intuitiva podemos asumir que mientras θ_Q captura el interés del usuario por el tema concreto expresado en su consulta, θ_D especifica el asunto del que trata el documento. Partiendo de esta hipótesis se puede aplicar la medida de divergencia KL, de forma que mida la similitud entre ambos modelos. Así, cuanto más parecidos sean ambos modelos más arriba en el ranking aparecerá dicho documento. Formalmente, dada una consulta Q y un documento D el valor asignado por la función de ranking a D vendrá dado por la siguiente expresión:

$$Score(D,Q) = -D_{KL}(\theta_Q||\theta_D) = -\sum_{w \in V} p(w|\theta_Q)\frac{p(w|\theta_Q)}{p(w|\theta_D)}$$

siendo V el vocabulario.

La estimación de θ_D puede realizarse de manera similar a como se desarrolla en el modelo de *Query Likelihood*, mientras que la forma más simple de estimar θ_Q se basa en la distribución empírica de las palabras en la consulta según la siguiente expresión:

$$\frac{frec_{w,Q}}{|Q|}$$

siendo $frec_{w,Q}$ la frecuencia de aparición del término w en la consulta Q y $|Q|$ el número total de palabras que contiene la consulta.

Relevance Model

Una tercera aproximación del uso de modelos de lenguaje que se centra en una estimación más adecuada de θ_Q, corresponde al modelo de recuperación denominado *Relevance Model*. Este enfoque se centra en incorporar información de relevancia al propio modelo. Así, θ_Q se estima en base a un conjunto de documentos considerados relevantes R, que suelen ser aquellos que aparecen más arriba en el ranking usando la aproximación *Query Likelihood*. De esta forma, la probabilidad de que una palabra sea generada por θ_Q vendrá dada por la media de las probabilidades de aparición en cada uno de los documentos de R, donde cada una de estas probabilidades parciales es multiplicada por el valor de *Query Likelihood* obtenido para dicho documento según la consulta original. Es decir, se utiliza el valor de *Query Likelihood* obtenido como factor que describe la importancia del documento D, mientras se toma como probabilidad de aparición de una palabra, la probabilidad promedio a partir de los distintos modelos de lenguaje del conjunto de documentos considerados relevantes. Esta aproximación se puede observar como el cálculo de una media ponderada de la probabilidad de aparición de la palabra en el subconjunto R.

Este modelo permite además incluir un factor de probabilidad a priori $p(\theta_D)$, que describe la probabilidad de que un documento sea relevante

debido a otras características, como por ejemplo su valor de *PageRank* o antigüedad.

Formalmente, la probabilidad de que un término haya sido generado por el modelo de lenguaje de la consulta, según la aproximación *Relevance Model*, se define con la siguiente expresión:

$$p(w|\theta_Q) = \sum_{\theta_D \in \Theta} p(w|\theta_D) p(\theta_D) \prod_{i=1}^{m} p(q_i|\theta_D)$$

donde Θ corresponde a los modelos de lenguaje de cada uno de los documentos de la colección; $\prod_{i=1}^{m} p(q_i|\theta_D)$ es el valor de *Query Likelihood* calculado usando la consulta original, siendo m la longitud de la consulta; $p(\theta_D)$ es el factor a priori y $p(w|\theta_D)$ es la probabilidad de que w haya sido generado por el modelo del lenguaje del documento. De esta forma aquellos términos que aparezcan en los documentos más relevantes según *Query Likelihood*, tendrán mayor probabilidad de haber sido generados por θ_Q.

Existe una segunda versión de *Relevance Model* denominada modelo 2 (Lavrenko y Croft (2001)) muy similar a ésta, pero en la que se calcula la probabilidad de aparición de una palabra usando documentos en los que aparecen w y al menos un término de la consulta original.

Una vez realizado un repaso por los distintos modelos de recuperación más relevantes para este trabajo, a continuación se profundiza en cómo estos modelos son evaluados bajo el denominado paradigma *Cranfield*.

2.2. *Cranfield* y calidad de respuesta

Dentro del campo de la recuperación de información y desde sus inicios, se ha mostrado como indispensable la existencia de una plataforma de pruebas suficientemente robusta, que permitiera medir la efectividad de los distintos modelos o métodos aplicados. Esta es la razón por la cual históricamente han aparecido una serie de colecciones que han ido incrementando con los años su tamaño y robustez (Voorhees (2002)).

En líneas generales una colección de pruebas debe consistir de tres componentes principales que permitan la evaluación de distintos sistemas:

- Un conjunto de documentos.
- Un conjunto de necesidades de información o *"topics"*, que se resuelvan al menos en parte, en el conjunto de documentos utilizado.
- Un conjunto de juicios de relevancia, que consisten en una asociación entre cada necesidad informativa y cada documento. Esta asociación se ha definido habitualmente en términos binarios, pudiendo tomar dos valores, *relevante* o *no relevante*, y por tanto indicando cuando un documento es relevante para una necesidad de información concreta.

De esta forma, es trivial conseguir un listado de aquellos documentos relevantes dada una necesidad de información, para posteriormente comparar esta lista con otra obtenida a través del método de recuperación que se pretenda evaluar.

Con el objeto de obtener una perspectiva histórica de la evolución de las colecciones a lo largo de los años, se presenta un breve repaso sobre las colecciones más destacadas en el campo de la recuperación de la información.

- *The Cranfield Collection*, fue la colección pionera y data de los años 50 del siglo pasado, contenía 1398 resúmenes de artículos que aparecieron en revistas especializadas en aerodinámica. Además se propusieron un total de 225 necesidades informativas sobre la colección. En la actualidad esta colección está en desuso ya que su pequeño tamaño la hace muy poco representativa.

- *Text Retrieval Conference*, más conocida como *TREC*, es una conferencia de carácter anual organizada por el *NIST*[2]. El objetivo de esta conferencia es el de construir una serie de plataformas de pruebas para la evaluación en el campo de la recuperación de la información. *TREC* se subdivide en distintas tareas o *"tracks"*, que se centran en un problema concreto, como pueden ser secciones especializadas en la Web, centradas en la recuperación de contenidos multimedia, orientadas a la empresa u otras. En relación a la recuperación de información *"ad-hoc"*, destacan las siguientes colecciones utilizadas a lo largo de los años en el curso de distintas tareas llevadas a cabo en *TREC*. La colección de documentos *TREC Vol. 4 +5* se empleó para la tarea de *"Robust"* en los años 2004 y 2005 y contiene alrededor de 528.000 documentos con un tamaño total de casi 2 Gigabytes (Voorhees (2006)). El conjunto de necesidades de información disponibles para dicha colección alcanza el número de 250 junto a sus correspondientes juicios de relevancia. Un primer paso para la evaluación de los sistemas en el entorno de la Web, fue la construcción de la colección *WT10g*, esta colección posee más de millón y medio de páginas Web (Chiang et al. (2005)), y existen disponibles para ella 100 necesidades de información. En los últimos años y para el entorno Web se ha construido la colección *GOV2*, que incluye más de 25 millones de páginas web (Clarke et al. (2005)) y para la que hay disponibles 150 necesidades de información.

- *CLEF* se puede considerar el *TREC* europeo. Debido a la diversidad lingüística existente en Europa se creó un congreso al estilo *TREC* de carácter anual, pero centrado sobre todo en recuperación de información multilingüe y en la construcción de recursos de evaluación para dicho objetivo.

[2]NIST: Instituto Nacional de Estándares y Tecnologías de los EEUU.

- Finalmente destacar otras colecciones como aquellas generadas por el *NTCIR*, que se centra en la construcción de colecciones para entorno multilingüe pero dentro del ámbito de los países del este asiático.

2.2.1. Medidas de evaluación

Una vez seleccionada la colección y las distintas necesidades de información sobre dicha colección, debemos definir qué medidas de evaluación serán las utilizadas. El objetivo de las medidas de evaluación es mostrar cuantitativamente la validez de las distintas técnicas propuestas, es decir, medir cómo de acertada es una aproximación para una colección específica. Las dos medidas básicas más extendidas son la precisión y la cobertura. Sobre estas medidas se ha construido todo un modelo de evaluación, que se describirá a continuación, y que se ha convertido en un estándar en el campo de la recuperación de la información.

Cobertura

Dada una consulta Q, realizada sobre una colección C, se denomina cobertura, al número de documentos relevantes recuperados en C en base a Q, dividido por el número total de documentos relevantes en C según Q, esto es:

$$\text{Cobertura} = \frac{\text{número de documentos relevantes recuperados}}{\text{número de documentos relevantes en la colección}}$$

De forma intuitiva se puede describir la cobertura como la capacidad de recuperar de la colección el máximo número de documentos relevantes. La cobertura se mueve entre los valores cero y uno, cero en el peor caso, es decir cuando no recuperamos ningún documento relevante, y uno cuando recuperamos todos los documentos relevantes existentes en nuestra colección.

En contra de la medida de cobertura se puede alegar que simplemente recuperando todos los documentos de nuestra colección, relevantes o no relevantes, conseguiríamos una cobertura igual a uno. Otra característica de la cobertura es que crece con el número de documentos recuperados, es decir dibuja una función creciente, que está correlacionada con el número de documentos recuperados.

La medida de cobertura es muy valorada en entornos en los que se hace imprescindible obtener todos los documentos que se consideran relevantes dada una necesidad informativa, y la omisión de algún documento no sea aceptable. Pensemos por ejemplo en un sistema de información jurídica o de historiales médicos en los que es imprescindible acceder a todos los documentos relacionados con un caso concreto.

En cambio en un entorno tan de moda en la actualidad como la Web, la medida de cobertura no es muy apreciada, ya que el número de documentos

relevantes para una necesidad de información puede ser muy elevado, del orden de cientos de miles o millones de documentos. Es en este ámbito donde más se valora la siguiente medida que se describe, la precisión.

Precisión

Dada una consulta Q, realizada sobre una colección C, se denomina precisión, al número de documentos relevantes recuperados en C en base a Q, dividido por el número total de documentos que se hayan recuperado, esto es:

$$\text{Precisión} = \frac{\text{número de documentos relevantes recuperados}}{\text{número total de documentos recuperados}}$$

Es decir, se calcula la fracción de documentos relevantes recuperados respecto al total de los recuperados. Por tanto, suministra una medida de lo "precisos" que somos al recuperar. El valor ideal será 1 y solo ocurrirá en caso de que todos los documentos recuperados sean relevantes. Cuantos más documentos no relevantes se recuperen menor será el valor de precisión. Un sistema ideal tendría cobertura y precisión igual a uno. Para conseguir este ideal respecto a una necesidad de información, sería necesario recuperar los documentos relevantes, y ninguno más.

En la realidad este ideal es difícilmente alcanzable. Es más, se da que la precisión y la cobertura son medidas que evolucionan inversamente, esto es, a mayor cobertura disminuirá la precisión y viceversa, como se puede observar en la Figura 2.2.

Precisión a *k*

Una extensión a la medida de precisión básica, sería la precisión a k ($P@k$). Esta medida indica la precisión obtenida en el conjunto de los primeros k documentos recuperados. Esta medida es muy valorada en entornos en los que se considera fundamental responder al usuario con una lista de documentos ordenada por relevancia. Un ejemplo clásico es la Web, donde es muy importante conseguir valores de precisión muy altos para k igual al número de resultados que muestra por defecto un buscador en la primera página. Es importante entender que en este entorno se prima devolver el mayor número de documentos relevantes para k pequeños. Obviamente, no es esperable que un usuario navegue hasta la página 5 o 6 de resultados de un buscador cualquiera, hasta encontrar la página que le interese. Una crítica a la utilización de esta medida, sería el hecho de que penaliza mucho las necesidades informativas con un número pequeño de documentos relevantes respecto a otra que tuviera un número elevado de documentos relevantes. Esto dificulta en gran medida, la comparación entre los valores obtenidos por dos necesidades de información distintas.

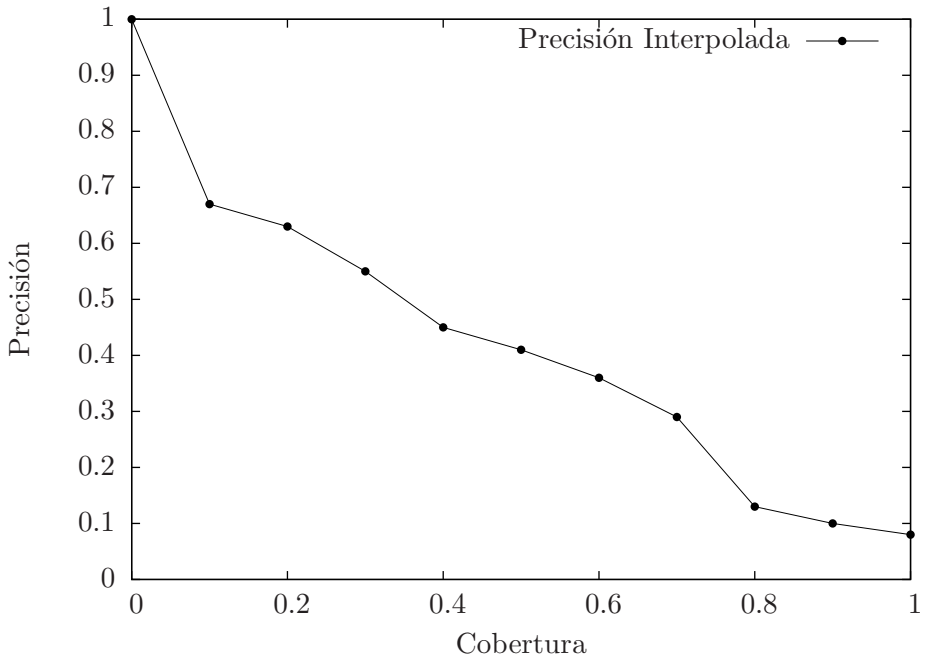

Figura 2.2: Gráfica típica de los valores obtenidos de Precisión respecto a los de Cobertura.

Medida F

El problema que surge con la utilización de precisión y cobertura, es consecuencia de manejar dos medidas diferentes y que tienden a mostrar un comportamiento inverso la una respecto a la otra. Esto ha llevado a construir funciones de evaluación más sofisticadas que permiten cuantificar una medida que combine ambas.

Una primera aproximación sería la utilización de la "medida-F" (van Rijsbergen (1979), pág. 134), que es una función que combina ambas medidas mediante una media armónica con pesos. Así la "medida-F" con los valores $\alpha = 1/2$ y $\beta = 1$, que establece la misma importancia para precisión y cobertura se expresaría como:

$$F_{\beta=1} = \frac{2PC}{P+C}$$

siendo P la precisión y C la cobertura.

MAP

MAP se ha convertido en uno de los estándares de facto para la comunidad de recuperación de la información en los últimos años, especialmente en

relación a *TREC* y a la recuperación de tipo *"ad-hoc"*. *MAP* cuyo significado es *"Mean Average Precision"*, calcula una media de la precisión hallada a distintos niveles de cobertura. Estos niveles de cobertura son tantos como documentos relevantes se recuperen. Así, se toma una medida de precisión cada vez que un sistema recupera un nuevo documento relevante.

En la literatura relacionada existen diversos trabajos que defienden la bondad de *MAP* para la evaluación de sistemas de recuperación de información, basándose en dos características principales. El hecho de que *MAP* es capaz de suministrar un valor único que combina precisión y cobertura de forma adecuada y su sensibilidad a pequeñas diferencias de rendimiento entre sistemas, lo que la convierte en una medida muy fiable (Buckley y Voorhees (2000); Sanderson y Zobel (2005)). Dicha sensibilidad tiene su soporte, en el hecho de que utiliza toda la información disponible de documentos juzgados como relevantes, en combinación con las diferencias en relación a la posición que ocupan dichos documentos en el ranking, al ser recuperados.

Formalmente, sea un conjunto de documentos relevantes:

$$Rel = [d_1, d_2, \ldots, d_n]$$

de cardinalidad n para una consulta q y sea R_j el conjunto de documentos previamente recuperados antes de recuperar el j-ésimo documento relevante del conjunto *Rel*. Se define la precisión media, *"Average Precision" (AP)*, como:

$$AP = \frac{1}{n} \sum_{j=1}^{n} Precision(R_j)$$

Es decir, se calcula la precisión media interpolada usando cada punto donde existe un cambio en la cobertura de los documentos recuperados.

Esta ecuación se puede extender fácilmente para el caso de querer obtener el promedio de la precisión media de un conjunto de consultas $Q = [q_1, q_2, \ldots, q_m]$

$$MAP = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{n} \sum_{j=1}^{n} Precisión(R_j)$$

Es decir simplemente se calcula la media aritmética entre los distintos valores de *AP* para cada necesidad informativa, obteniendo de esta forma el valor de *MAP*.

Estas son algunas de las medidas más habituales aplicadas en la evaluación cuantitativa de la calidad de un conjunto de documentos devueltos a partir de una consulta. De manera más específica, *MAP* es la medida de evaluación que se utiliza mayoritariamente en esta tesis, ya que se considera un estándar en la actualidad, para evaluar la validez de las distintas técnicas propuestas.

2.3. Refinamiento de consultas

La calidad de la respuesta de un sistema de recuperación de información, depende en gran medida del acierto del usuario a la hora de expresar su necesidad de información. En el caso de que la consulta no describa acertadamente la idea o concepto que se desea expresar, cualquier sistema se verá incapaz de recuperar aquellos documentos que satisfagan las expectativas del usuario.

Este problema ocurre con bastante frecuencia, ya que es común que los usuarios no sean capaces de expresar con exactitud aquella información que desean recuperar, bien porque no sepan exactamente que están buscando o simplemente porque el vocabulario que utilizan no se corresponde con aquél que aparece en los documentos más relevantes. Sin embargo, si a ese mismo usuario se le suministra una muestra de documentos potencialmente relevantes, le resultará relativamente fácil reconocer aquellos que son de utilidad para él. Este hecho ha provocado el desarrollo de diversos modelos, que aprovechándose de esta característica, son capaces de incorporar información de relevancia a la consulta original, con el objetivo de mejorar la calidad de la respuesta que se devolverá al usuario. Este tipo de metodologías se denominan de forma genérica técnicas de Refinamiento de la Consulta o según la terminología original *Relevance Feedback*.

Aproximaciones al refinamiento de consultas

La muestra de documentos relevantes se puede obtener bien de forma explícita, solicitando al usuario que indique que documentos considera relevantes, o bien de forma automática es decir tratando de inferir en base a la consulta original un subconjunto de documentos relevantes.

Existen diversos métodos que permiten modificar la consulta original a partir de la muestra obtenida. En líneas generales, estas técnicas consisten en seleccionar de un conjunto de términos candidatos, que habrán sido extraídos de los documentos considerados relevantes, aquellos que permitan recuperar un mayor número de documentos adecuados para la necesidad de información original.

Existen métodos específicos para cada modelo teórico de recuperación que actúan a modo de extensiones para estos. Así, para el modelo de espacio vectorial y aquellos basados en el esquema de pesado *TF-IDF* nos encontramos con el método de Rocchio (Rocchio (1971)). En relación con el modelo probabilístico destacan las distintas propuestas realizadas por Robertson (Robertson (1986, 1991)) mientras que para el caso de los modelos de lenguaje destaca la aproximación *Relevance Models* que aparece en Lavrenko y Croft (2001). Un análisis más detallado sobre los distintos métodos de refinamiento de consultas escapan del ámbito de esta tesis, pero una descripción en gran detalle puede encontrarse en Ruthven y Lalmas (2003).

Este tipo de metodologías se ha mostrado bastante efectiva cuando la información de relevancia parte del usuario que ha expresado la necesidad de información. En cambio, la utilización de métodos automáticos no ha sido tan exitosa. Típicamente el problema que ocurre con la aplicación de los métodos automáticos, es la dificultad para inferir qué es relevante para el usuario a la hora de obtener la muestra necesaria de documentos relevantes. La calidad del resultado obtenido tras la aplicación de estos métodos depende en gran medida del acierto a la hora de seleccionar documentos realmente relevantes, hecho este que depende principalmente de la calidad de la consulta original. Por tanto, se hace difícil prever cuando una consulta modificada mediante métodos automáticos obtendrá una mejor respuesta respecto a la consulta original especificada por el usuario.

Stemming: Un caso de uso de refinamiento de consultas

Estudios relacionados con las limitaciones que presentan los métodos automáticos de refinamiento, y que en cierta manera están entre las bases de algunas de las ideas que se desarrollarán en esta tesis, aparecen en los siguientes trabajos (Araujo et al. (2010), Pérez-Agüera et al. (2008), Araujo y Pérez-Agüera (2008)). En estos, se presentaba un método de optimización basado en heurísticas, que ayudaba a seleccionar adecuadamente aquellos conjuntos de términos que incrementaban la calidad de la respuesta respecto a la consulta original, para el caso concreto del *stemming*.

El *stemming* es un preproceso a la indexación de contenidos realizado sobre el texto que aparece en los documentos a indexar. Esta técnica se utiliza habitualmente con el fin de reducir distintas palabras a sus raíces morfológicas comunes, por ejemplo eliminando las distintas variaciones para una misma palabra como plural, gerundio o género (Baeza-Yates y Ribeiro-Neto (1999), cap. 7). La intuición detrás de la utilidad del *stemming* se basa en que es capaz de agrupar conceptos similares expresados con palabras distintas que comparten la misma raíz. Aunque la intuición anterior parece perfectamente plausible, no existe una clara posición en la literatura acerca de si la aplicación de procesos de *stemming*, mejora la calidad de las respuestas como aparece reflejado en el trabajo Fox (1992). Es relativamente fácil encontrar ejemplos en los que la aplicación del *stemming* dañaría claramente el rendimiento de una consulta. Por ejemplo, en el caso de las palabras *beber* y *bebes* ambas comparten la misma raíz aunque por el contrario los conceptos a los que se refieren son radicalmente distintos.

El proceso de *stemming*, que se suele realizar de forma previa a la construcción del índice de documentos, también puede realizarse en tiempo de consulta de forma similar al refinamiento de consultas, es decir, añadiendo aquellas palabras que compartiendo raíz con los términos de la consulta original se espera que puedan mejorar la calidad de la respuesta. Este era uno de los objetivos propuestos en los trabajos mencionados previamente, ana-

lizar el comportamiento del proceso de *stemming* en tiempo de indexación, respecto a su aplicación en el momento de búsqueda.

En Araujo y Pérez-Agüera (2008), se comprobaba experimentalmente que la introducción del conjunto total de palabras que comparten la misma raíz a la consulta original provocaba un descenso importante en la calidad de los resultados obtenidos respecto a la utilización de la consulta original. Para mejorar dichos resultados, se proponía la aplicación de un método de optimización basado en heurísticas que permitiera, en un tiempo razonable, seleccionar aquel subconjunto de términos más adecuado para la consulta original. El principal problema para seleccionar el subconjunto de términos óptimo, es el ingente número de posibles soluciones que impide una evaluación exhaustiva de la calidad de cada una de las alternativas. Así, se opta por aplicar un método de optimización basado en algoritmos genéticos, junto con una función de ajuste que fuera capaz de estimar la adecuación de un subconjunto específico de términos para realizar la expansión.

La función de ajuste propuesta se basaba en el cálculo del coseno entre la consulta expandida y el documento recuperado en primer lugar con dicha consulta. De esta forma, se esperaba que un subconjunto de términos candidatos fuera más adecuado cuanto más cercano a 1 fuera el valor del coseno calculado entre la consulta modificada y el documento recuperado en primer lugar con dicha consulta. La idea de usar como función de ajuste el coseno se basa en la intuición de que era más probable que aquellos términos conceptualmente más cercanos co-aparecieran en el mismo documento, lo que por tanto eleva el valor del cómputo del coseno; simultáneamente aquellas palabras que compartiendo raíz describieran conceptos alejados aparecerían juntos en un menor número de casos, como en el ejemplo anterior en relación a los términos *beber* y *bebe*.

Una de las principales conclusiones que se extrajeron a partir de este trabajo, fue comprobar experimentalmente que el valor obtenido al calcular el coseno está relacionado en cierto grado con la idoneidad de un documento respecto a la consulta con la que fue recuperado. Esta primera observación, se encuentra entre los fundamentos del trabajo que se pretende realizar en esta tesis. Esto es, se desea comprobar si el cálculo de ciertos estadísticos sobre los valores asignados por las distintas funciones de ranking, permiten predecir la calidad de la respuesta dada por un motor de búsqueda. Esta hipótesis se englobaría dentro del campo de la predicción del rendimiento de consultas o *Query Performance Prediction (QPP)*[3], siendo éste un campo de investigación en pleno desarrollo dentro de la RI y que se describe en detalle en la siguiente sección.

[3]En la literatura actual es común encontrar como sinónimos de *Query Performance Prediction* a QPP, predicción del rendimiento de consultas o predicción de la calidad de consultas, por lo que en este trabajo se utilizarán indistintamente.

2.4. Predicción del rendimiento de consultas *(QPP)*

En los últimos años se ha observado un creciente interés en el campo de la predicción de la calidad de consultas por parte de la comunidad de recuperación de información. Prueba de este hecho, es el elevado número de publicaciones relacionadas que han ido apareciendo en algunas de las conferencias de mayor impacto en el área.

Uno de los hitos dentro del campo de la predicción de la calidad de las consultas fue la publicación de los trabajos realizados por Cronen-Townsend et al. (2002). En este artículo se introducía un nuevo método de predicción denominado *Clarity Score*, que mostró la posibilidad de realizar predicciones con cierto grado de éxito. Además, en esta publicación se plasmó lo que se convertiría en el marco de investigación clásico en la investigación de este tipo de técnicas. Posteriormente, los trabajos desarrollados por Cronen-Townsen junto a Bruce Croft dieron lugar a la publicación de la tesis doctoral Zhou (2007), donde se resumía la investigación realizada en este campo por los autores.

Los distintos métodos de predicción del rendimiento de consultas tratan de estimar, si para una consulta específica expresada por un usuario, un sistema será capaz de recuperar un conjunto de documentos que sean relevantes. Es decir, el principal objetivo de la predicción es estimar la calidad del conjunto de documentos devueltos a un usuario en ausencia de juicios de relevancia, o lo que sería equivalente predecir los valores de calidad obtenidos usando medidas de evaluación estándar.

El rendimiento de un método de predicción, y por tanto de la calidad de sus predicciones, se evalúa en términos de correlación entre las estimaciones realizadas por el método de predicción y la medida de calidad que se considera objetivo. De hecho, la evaluación de un método de predicción tiene una fuerte dependencia con la medida de calidad seleccionada respecto a la cual, se desea que las estimaciones muestren la máxima correlación. En la gran mayoría de los trabajos relacionados, la medida de calidad a predecir suele corresponder con la precisión media AP.

En la Figura 2.3 se puede observar gráficamente los componentes básicos que forman un esquema genérico de predicción, donde a partir de una consulta se obtienen las predicciones sobre los valores de calidad objetivo que se calculan usando los juicios de relevancia.

Formalmente, dado un conjunto de N consultas Q, los valores de calidad que obtiene cada una de las consultas según una función objetivo $Obj_{q1}, Obj_{q2}, \ldots, Obj_{qN}$, y las estimaciones obtenidas con un método de predicción cualquiera $Pr_{q1}, Pr_{q2}, \ldots, Pr_{qN}$, la calidad de las estimaciones se evalúa según la siguiente expresión:

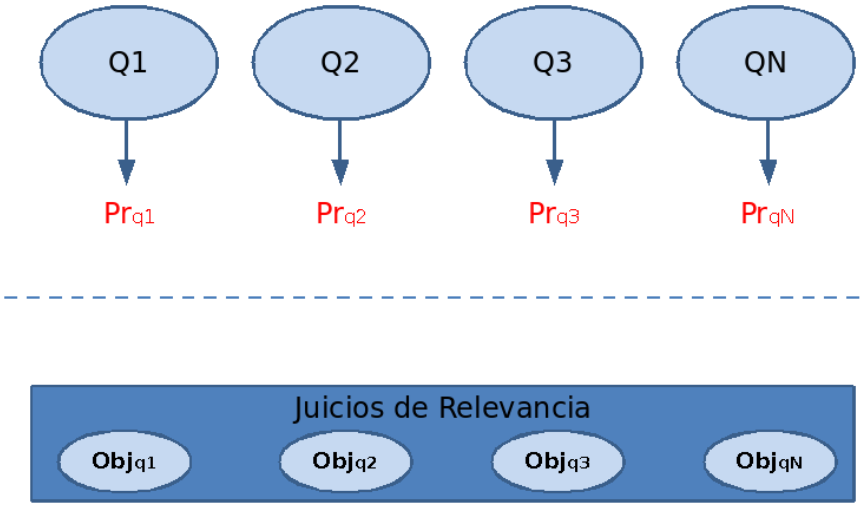

Figura 2.3: Estimaciones vs. función de calidad objetivo.

$$\forall_{q_i \in Q} corr(Obj_{qi}, P_{qi})$$

donde *corr* corresponde con la medida de correlación que se desee utilizar.

La habilidad principal que añade la inclusión de las diversas técnicas de predicción en un sistema de RI, es que el sistema de recuperación, en base a la información de calidad estimada, puede llevar a cabo medidas específicas que mejoren los procesos de búsqueda. Por ejemplo, en el caso de que la respuesta obtenida con la consulta original no tuviera calidad suficiente, permitiría indicar al usuario que dadas las características de su consulta no es posible suministrarle documentos de calidad, por lo que podría sugerirle modificar la consulta de forma que ésta fuera más adecuada.

Existen dos tipos de aproximaciones principales a la hora de realizar la predicción. Así, se distingue entre aquellos métodos que realizan la predicción previamente a la recuperación, denominados métodos *Pre-Retrieval*, y aquellos otros que la realizan posteriormente a la recuperación, denominados métodos *Post-Retrieval*.

La principal diferencia entre ambas aproximaciones radica en la utilización del conjunto de documentos devueltos a partir de una consulta para realizar la predicción. Así, mientras los primeros no utilizan esta información, los métodos *Post-Retrieval* utilizan la lista de resultados para realizar sus predicciones. Como consecuencia de este hecho se considera que en general los métodos *Post-Retrieval* realizan predicciones más acertadas ya que cuen-

tan con mayor información, los documentos recuperados en primer término, aunque este hecho implica un coste computacional más elevado.

2.4.1. Métodos de predicción *Pre-Retrieval*

Este tipo de métodos realizan sus predicciones independientemente del conjunto de documentos que devolvería un SRI dada una consulta y como consecuencia son menos dependientes del modelo de recuperación aplicado. Este hecho implica que su única fuente de información, a la hora de predecir, son los propios términos que componen la consulta, y por tanto los distintos estadísticos de estos términos calculados en base a la colección de documentos u otras fuentes externas.

Este tipo de aproximaciones se centran en la estimación de la dificultad de una consulta de forma general, más allá del rendimiento de dicha consulta en un sistema particular. Aunque, puesto que la evaluación de un método de predicción se realiza respecto a la calidad obtenida por un sistema concreto, las aproximaciones *Pre-Retrieval* también muestran cierta dependencia respecto a las características del sistema, a partir del cual se obtiene el valor de la medida objetivo.

Existe un número muy amplio de métodos de predicción de tipo *Pre-Retrieval*. En especial destacan los trabajos publicados en He y Ounis (2004) y He y Ounis (2006). En ambos trabajos los autores describen la especificidad de una consulta como una de las característica principales a la hora de estimar la calidad de la respuesta dada a dicha consulta. Así, los autores argumentan que las consultas más específicas obtendrán una respuesta de mayor calidad respecto a aquellas más generales o ambiguas. En base a esta suposición, los autores presentan una familia de métodos de predicción basados en la especificidad de la consulta y en como estimar dicha característica a partir de los términos de ésta.

La aproximación más sencilla se basa en la longitud de las consultas, partiendo de la intuición de que las consultas que cuentan con más palabras son más específicas. Así, esta primera aproximación a la predicción se basa en el número de términos que contiene una consulta, obviando las denominadas palabras vacías.

Basados en *IDF* o *ICTF*

Otro conjunto de métodos de predicción propuesto por los mismos autores, corresponde a los métodos basados en la frecuencia inversa de documento (*IDF*). Dicho estadístico es un componente básico en la mayoría de las funciones de ranking y mide lo discriminante que es un término para una colección. Así, si un término aparece en un gran número de los documentos de la colección se considera poco discriminante, ya que al incorporarlo a una consulta recuperará un número muy elevado de documentos. Los autores

destacan dicho estadístico como una característica importante que permite estimar la dificultad de una consulta, basándose en lo expuesto previamente por Pirkola y Järvelin (2001). Utilizando esta misma idea los autores definen diferentes métodos de predicción basados en *IDF*. La intuición común a todas estas propuestas es estimar "como de específica es la consulta". Dada una consulta Q que contenga los términos $t_1, t_2, ..., t_n$, los métodos de predicción propuestos se definen formalmente como a continuación:

- El valor de predicción viene dado por el mayor valor de *IDF* encontrado en los términos de la consulta:

$$IDF_{max} = argmax_{t_i \in Q}(IDF(t_i))$$

- El valor de predicción viene dado por el menor valor de *IDF* encontrado en los términos de la consulta, por lo que el valor de predicción debe ser inversamente proporcional a la calidad de la respuesta:

$$IDF_{min} = argmin_{t_i \in Q}(IDF(t_i))$$

- El valor de predicción viene dado por el promedio de los valores de *IDF* de la consulta:

$$IDF_{avg} = \frac{1}{n}\sum_{i=1}^{n} IDF(t_i)$$

Siguiendo con la idea de la especificidad de la consulta, los autores definen dos nuevos métodos de predicción cuyo objetivo es estimar la distribución de los valores de *IDF* que toman los términos de la consulta. Así, dicha distribución se estima basándose en la desviación estándar o en el ratio encontrado entre el máximo y mínimo valor de *IDF*, como aparece a continuación:

- Basado en la desviación:

$$\gamma_1 = \sigma_{IDF} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(IDF_{t_i} - \overline{IDF_t})^2}$$

  siendo $\overline{IDF_t}$ el valor promedio de *IDF* para todos los términos de la consulta Q.

- Basado en el ratio entre el máximo y el mínimo IDF:

$$\gamma_2 = \frac{IDF_{max}}{IDF_{min}}$$

Los métodos de predicción anteriores, γ_1 y γ_2, tienen como principal limitación que para consultas de un solo término el valor que se obtiene permanece constante. Para el caso de la desviación será equivalente a cero, ya que la desviación de un único elemento es siempre cero, y para el caso del ratio entre máximo y mínimo será uno, ya que para consultas de un solo término $IDF_{max} = IDF_{min}$.

Es interesante destacar que aunque los autores de este trabajo definen *IDF* como

$$IDF(t) = \frac{log_2(N + 0{,}5)/N_t}{log_2(N + 1)}$$

donde N_t es el número de documentos donde aparece el término t y N es el número total de documentos en la colección, trabajos desarrollados posteriormente (Hauff et al. (2008a)) muestran similares resultados a los obtenidos por los autores con el uso de la formulación clásica de *IDF*, esto es:

$$IDF(t) = log\left(\frac{N}{N_t}\right)$$

Otra familia de métodos de predicción que se basa igualmente en la especificidad de la consulta son aquellos que en lugar de utilizar *IDF* se decantan por el uso de *ICTF* (He y Ounis (2004) y He y Ounis (2006)).

ICTF corresponde con las siglas en inglés de *Inverse Collection Term Frequency*, o en español frecuencia inversa de un término en la colección. La idea que subyace detrás de *ICTF* es equivalente a la aplicada en *IDF* ya que ambas medidas representan de forma similar la capacidad de recuperación de un término específico. Así, mientras *IDF* utiliza el número de documentos donde aparece el término, *ICTF* utiliza la frecuencia total en la colección.

El valor $ICTF$ para el término t se define de la siguiente forma:

$$ICTF_t = \frac{longitud(C)}{frec(t)}$$

esto es, el número de términos que aparecen en el conjunto de documentos de la colección, dividido por el número de veces que aparece el término t en la misma colección. Así, de forma similar a como se definía el método de predicción IDF_{avg}, los autores definen un nuevo método de predicción calculando el promedio de los valores de *ICTF* para los distintos términos de la consulta, quedando la expresión como se observa en la siguiente ecuación:

$$AvICTF = \frac{1}{n}\sum_{i=1}^{n} ICTF(t_i)$$

Simplified Clarity Score

Una nueva aproximación que se fundamenta en la misma intuición que hay detrás de *Clarity Score* aparece en He y Ounis (2004). Los autores denominan a este método *Simplified Clarity Score* (*SCS*), y es una simplificación

del modelo de *Clarity Score* para el caso *Pre-Retrieval*. Dicho método se fundamenta en las diferencias que pueden ser encontradas entre el modelo de lenguaje de la consulta respecto al de la colección. La diferencia entre ambos modelos de lenguaje se computa con el cálculo de la divergencia *Kullback-Leibler*, de forma similar a como se realiza en *Clarity Score*.

La principal diferencia con *Clarity Score* aparece a la hora de estimar el modelo de lenguaje de la consulta. Aquí dicha estimación se computa utilizando solamente el conjunto de términos que aparecen en la consulta, de forma que el modelo de lenguaje para una consulta q que contenga el término q_i se define como:

$$Pml(q_i|q) = \frac{frec(q_i)}{n}$$

donde n es el número total de términos que aparecen en la consulta; mientras que el modelo de la colección se estima como es habitual:

$$Pml(q_i|C) = \frac{frec(q_i)}{longitud}$$

siendo *longitud* el número de términos de la colección. Formalmente *SCS* se define:

$$SCS = \sum_{i=1}^{n} Pml(q_i|Q) log_2 \frac{Pml(q_i|Q)}{Pml(q_i|C)}$$

para una consulta Q que contenga n términos.

La idea que subyace a este método es estimar la coincidencia entre el modelo de lenguaje de la consulta y el de la colección, de tal forma que si dicha diferencia es mínima, se considera que la consulta es poco discriminante ya que se ajusta al total de la colección. Por el contrario, en el caso de que dicho valor de divergencia sea elevado implica que la consulta es específica y que por tanto la calidad de la respuesta será elevada. Es importante destacar que una alta divergencia también podría implicar que el asunto del que trata la consulta carezca de representación a lo largo de la colección, lo que nos llevaría a obtener una pobre respuesta para dicha consulta.

Alcance de la consulta

Finalmente los autores proponen el método *QueryScope* centrado en medir el grado de alcance de la consulta, es decir la representatividad de la consulta en la colección.

QueryScope utiliza como estimador el número de documentos que contienen al menos alguno de los términos que aparecen en la consulta original. De manera intuitiva podemos suponer que si encontramos un gran número de documentos que contengan algún término de la consulta, ésta será menos específica y por tanto el valor de predicción que describa la calidad de

la consulta deberá ser menor. Así *QueryScope* se define según la siguiente expresión:

$$QueryScope = -log\frac{N_q}{N}$$

donde N_q es el número de documentos que contienen al menos un términos de la consulta y N es el número total de documentos que aparecen en la colección.

Similitud entre la consulta y la colección

Otra familia de métodos de predicción que ha mostrado cierto potencial aparece en los trabajos de Zhao et al. (2008), donde se presentan dos enfoques diferentes de métodos de predicción. En primer término aparecen aquellos basados en lo que los autores denominan *Collection Query Similarity* o *SCQ*, que corresponde con una combinación de los valores de *IDF* e *ICTF* obtenidos a partir de los términos que aparecen en la consulta. Como se ha observado anteriormente la obtención de valores elevados para cualquiera de ambos estadísticos puede implicar una mayor calidad en la respuesta a obtener, por lo que de forma natural la combinación de ambos valores deberá seguir un comportamiento similar. A partir de esta idea los autores definen *SCQ* como aparece a continuación:

$$SCQ = (1 + ln(cf(q_i)))ln\left(1 + \frac{N}{df(q_i)}\right)$$

donde $cf(q_i)$ es el número de veces que aparece el término de la consulta q_i en la colección, $df(q_i)$ es el número de documentos en los que aparece el término de la consulta q_i y N es el número de documentos en la colección.

A partir de la expresión anterior los autores definen los siguientes métodos de predicción en base a distintas variaciones calculadas sobre el valor *SCQ* para cada uno de los términos:

- $SumSCQ$: La suma de los valores de *SCQ* para todos los términos de la consulta original Q.
- $MaxSCQ$: El valor máximo de *SCQ* sobre los valores de Q.
- $AvgSCQ$: El promedio de los valores de *SCQ* para todos los términos de la consulta.

Como segunda aproximación en el mismo trabajo (Zhao et al. (2008)), los autores proponen un enfoque basado en la variabilidad de los términos de la consulta original. Así, en esta aproximación los autores sugieren la variabilidad de los pesos asignados por la función de ranking, a cada uno de los términos de la consulta como indicio de la calidad de ésta.

La intuición en la que se fundamenta esta nueva aproximación se basa en observar si el peso asignado a un término único de la consulta q_i es equivalente o muy similar en una gran mayoría de los documentos de la colección en los que dicho termino aparece. En este caso se puede asumir la no existencia de evidencias que permitan decidir cuál de estos documentos es más relevante dado q_i. Los autores deciden utilizar un esquema de pesado basado en *TF-IDF*, de forma que a un término q_i se le asigna el valor $weight(q_i, d)$ para el documento d.

A partir de esta idea los autores proponen dos expresiones distintas que se fundamentan en el análisis de la variabilidad. Así, por un lado proponen el sumatorio de las desviaciones $SumVAR$, y por otro $AvVAR$ o el promedio de las desviaciones de cada término. Por tanto $AvVAR$ es equivalente al valor obtenido con $SumVAR$ dividido por el número de términos n de la consulta, esto es $\frac{1}{n}SumVAR$, donde $SumVAR$ se define como:

$$SumVAR = \sum_{i=1}^{n} \sqrt{\frac{1}{df(q_i)} \sum_{q_i \in d} (weight(q_i, d) - \overline{weight}_{q_i})^2}$$

siendo $\overline{weight}_{q_i}$ el promedio de los pesos asignados a q_i en los documentos donde aparece q_i y n el número de términos que componen la consulta.

Pointwise Mutual Information

Un enfoque distinto basado en la co-aparición de los términos de la consulta, más allá de la aparición de estos de forma individual aparece en Hauff et al. (2008a). Aquí se proponen dos métodos basados en el valor de *PMI Pointwise Mutual Information*. Esta medida describe el valor de discrepancia entre dos variables aleatorias comparando la probabilidad de que ambas co-ocurran de forma aleatoria $P(a) \cdot P(b)$, respecto a las veces que realmente co-ocurren $P(a, b)$ en una muestra. Así, si ambas variables, en este caso si ambos términos, co-ocurren de forma habitual, es decir con una probabilidad mayor de la puramente aleatoria, se infiere que ambos términos están relacionados y que por tanto la consulta tiene un mayor grado de especificidad que en el caso en que los términos no tuvieran relación entre sí. Por tanto, dada una consulta Q de la cuál se extraen dos términos q_i y q_j, se define su valor de PMI como:

$$PMI = log\left(\frac{P(q_i, q_j)}{P(q_i)P(q_j)}\right)$$

En Hauff et al. (2008a) se proponen dos medidas basadas en el valor de PMI obtenido para cada par de términos posibles de la consulta:

- $AvPMI$: el promedio de PMI obtenido para cada par de términos.

- $MaxPMI$: el máximo PMI encontrado entre todos los pares de la consulta original.

Otras aproximaciones

Se debe destacar la aproximación realizada por Mothe y Tanguy (2005) y que se aleja de los enfoques estadísticos presentados en esta sección. En este caso la predicción se basa en relaciones puramente semánticas o lingüísticas de los términos de la consulta. Aquí los autores presentan una familia de métodos de predicción basados en la distancia semántica de los términos de la consulta, donde se usa la distancia entre los términos calculada a partir de *Wordnet*.

Finalmente He et al. (2008), propone un método basado en el grado de solapamiento de aquellos grupos de documentos que se obtienen a partir de la aplicación de un proceso de *clustering* aglomerativo con los términos de la consulta. Aunque los resultados que se obtuvieron con este método fueron bastante prometedores, este tipo de aproximaciones basadas en el agrupamiento de documentos no ha continuado siendo explorada, ya que su aplicación a entornos reales es prácticamente inviable. El problema principal de estos enfoques es el coste computacional derivado de la realización de *clustering* aglomerativo sobre un conjunto de documentos muy amplio, junto con la tendencia que impera actualmente en el campo de la RI en relación al uso de colecciones de test de gran tamaño.

2.4.2. Métodos de predicción *Post-Retrieval*

Los métodos de predicción *Post-Retrieval* corresponden con la otra aproximación desarrollada para la predicción de la calidad de consultas. Se considera que esta aproximación tiene como principal ventaja, la posibilidad de utilizar la información del conjunto de documentos recuperados por una función de ranking a partir de la consulta. De esta forma es posible obtener mejores resultados respecto a los métodos *Pre-Retrieval* usando dicha información extra.

Clarity Score

La propuesta más exitosa y de mayor trascendencia dentro de este grupo de métodos de predicción corresponde a la realizada por Cronen-Townsend et al. (2002) en su método *Clarity Score*. La intuición detrás de este método es la de medir la ambigüedad de la consulta en relación a la colección de documentos. Esta aproximación se apoya en gran medida en los trabajos desarrollados anteriormente por los mismos autores, en relación con la aplicación de modelos de lenguaje para recuperación de información.

Para calcular el grado de ambigüedad entre la consulta y el corpus de documentos, los autores proponen usar la divergencia *Kullback-Leiber*. Esta

medida estima la divergencia entre los modelos de lenguaje correspondientes al conjunto de documentos devueltos por el motor de búsqueda utilizando la consulta original y el total de la colección. De esta forma los autores sugieren que cuanto mayor sea el valor de divergencia, mayor será la calidad de la respuesta respecto a dicha consulta.

Este razonamiento se basa en la suposición de que el conjunto total de documentos de la colección debe cubrir un amplio rango de temas, mientras que el conjunto de documentos devueltos, al menos el subconjunto más relevante, debería tratar de un tema específico estrechamente relacionado con la consulta original. De esta forma, si la divergencia es poco significativa, se puede concluir que la respuesta es poco precisa debido al amplio conjunto de temas que están representados en los documentos devueltos, mientras que si por el contrario dicha divergencia es elevada, se debe a que dicho conjunto de documentos trata sobre un único tema que coincidiría con el concepto expresado en la consulta.

Partiendo de la idea anterior los autores proponen estimar los distintos modelos de lenguaje de forma muy similar a la especificada en la aproximación *Relevance Model* en Lavrenko y Croft (2001). De esta forma el valor de *Clarity* se define como:

$$Clarity(Q) = \sum_{w \in V} P(w|Q) log_2 \frac{P(w|Q)}{P(w|C)}$$

siendo V el vocabulario en la colección. La estimación de $P(w|Q)$ se realiza a partir del conjunto R de los documentos que contengan al menos un término que aparezca en la consulta original, así:

$$P(w|Q) = \sum_{d \in R} P(w|D) P(D|Q)$$

donde $P(D|Q)$ se define como:

$$P(D|Q) = \prod_{q \in Q} p(q|D)$$

esto es, el producto de las probabilidades de que los términos de la consulta hayan sido generados por el documento D. Este factor asigna un peso o importancia a cada uno de los documentos en R, de forma que aquellos documentos que contengan una mayor frecuencia de los términos de la consulta obtendrán un valor mayor.

Por otra parte se define $P(w|D)$ utilizando el suavizado *Jelinek-Mercer* como aparece a continuación:

$$P(w|D) = \lambda P(w|D) + (1 - \lambda) P(w|C)$$

donde $P(w|D)$ es la frecuencia de aparición de w en el documento D y $P(w|C)$ es la frecuencia de aparición de w en el total de la colección.

Métodos basados en la perturbación de la consulta o documentos

Otra aproximación que ha sido explotada en distintos trabajos es aquella basada en la perturbación de la consulta original, expresada por el usuario según su necesidad de información. Así, esta aproximación consiste es realizar pequeñas modificaciones sobre la consulta original y observar las diferencias que se obtienen usando la consulta original respecto a la modificada, en cuanto a los documentos que se recuperan con ambas aproximaciones.

La intuición detrás de este enfoque se fundamenta en la idea de que aquellas consultas que al ser modificadas producen pequeñas variaciones en el conjunto de documentos recuperados, contienen una descripción muy específica del asunto del que tratan y que por tanto no se desvían de dicho asunto con facilidad. Por el contrario, en el caso de que la consulta se desviase en gran medida del asunto inicial, implicaría que la consulta no está perfectamente definida y que por tanto algunos de los documentos que se recuperan con ella no son relevantes de acuerdo a la necesidad de información.

Dentro de este enfoque destacan los trabajos realizados por Zhou y Croft (2007), donde modelan la hipótesis anterior en el método denominado *Query Feedback*. En este método se define la capacidad de una consulta para recuperar documentos relevantes como un problema de ruido en un canal de comunicación. Partiendo de la consulta original Q y de la lista de resultados L obtenida a partir de Q, se genera una nueva consulta Q' en base a los documentos que aparecen en L.

Una vez obtenida Q' se lanza esta nueva consulta modificada al motor de búsqueda, consiguiendo una nueva lista de resultados L'. A partir de ambas listas de resultados L y L' se utiliza el grado de solapamiento entre ellas como método de predicción. Un mayor solapamiento implica una mayor definición del asunto específico de la consulta, mientras que solapamientos bajos permiten intuir que la consulta derivada Q' está desviada en cierto grado de la temática original. En este mismo trabajo y basándose en la misma idea acerca de la perturbación de la consulta, los autores proponen el método *Weighted Information Gain*. En este caso se derivan distintas consultas a partir de la consulta original de forma equivalente al método anterior. A continuación, se mide la diferencia, en términos de probabilidad, de que las consultas derivadas aparezcan en el conjunto de documentos obtenidos con la consulta original, respecto a su presencia en el total de la colección. Así, se infiere que una consulta tendrá un mayor grado de calidad cuanto mayor sea la probabilidad de que las nuevas consultas aparezcan en los resultados originales, respecto a la probabilidad de que aparezcan en el total de la colección.

De forma similar, Yom-Tov et al. (2005) utilizan el grado de solapamiento entre los resultados obtenidos a partir de diferentes consultas derivadas de la consulta original. Pero a diferencia de *Query Feedback*, las consultas derivadas se construyen utilizando los distintos términos que aparecen en

la consulta original Q. Así, a partir de una consulta Q con m términos se obtienen m consultas derivadas $(Q_1, Q_2, \ldots Q_m)$, donde cada una de estas nuevas consultas contiene exactamente un único término de la consulta Q. De nuevo, en el caso de que se observe un alto grado de solapamiento entre los documentos recuperados por las distintas consultas derivadas y la original, se concluye que la consulta original está bien construida y que por tanto obtendrá un alto rendimiento.

Como extensión al trabajo anterior, Vinay et al. (2006) proponen modificar la consulta original mediante la asignación de pesos que describan la importancia de cada término en la consulta. Los pesos que se asignan a cada uno de los m términos de la consulta original Q, marcan la importancia relativa de estos términos respecto a la consulta. De nuevo el valor de predicción se obtendrá midiendo el grado de solapamiento entre las consultas derivadas y la original en relación a los documentos recuperados.

Una heurística estrechamente relacionada con la perturbación de la consulta, que aparece en Zhou y Croft (2006), consiste en modificar los documentos devueltos a partir de la consulta original en lugar de la propia consulta. Así, a partir de una consulta Q se obtiene un conjunto de documentos R y se modifican aleatoriamente algunos de los documentos de R construyendo un nuevo subconjunto R'. A continuación, se reordenan los documentos en R' según la consulta Q, estimándose la calidad de Q a partir de las diferencias observadas entre el orden de R respecto al de R'.

Partiendo del mismo enfoque de introducir cierto grado de perturbación en los documentos recuperados, Vinay et al. (2006) proponen utilizar como nueva consulta el primer documento recuperado a partir de Q modificándole levemente. Es inmediato comprobar que la utilización de un documento $d \in C$ como consulta sobre la colección C provocaría que este mismo documento fuera devuelto en primer lugar del ranking. De forma similar, si en lugar de utilizar como consulta el documento d utilizáramos este mismo documento con ciertas modificaciones d', es posible que d ya no apareciera en el primer lugar del ranking. Por tanto, en el caso de que la posición de d en el ranking devuelto usando como consulta d' descienda de forma significativa, se puede estimar que el grado de indefinición de la consulta Q es bastante elevado y como conclusión será bajo el rendimiento esperado de la consulta original.

Finalmente dentro de los métodos basados en perturbación de los documentos destaca la propuesta realizada por Diaz (2007). En este caso no son los documentos los modificados sino los valores de relevancia asignados por la función de ranking. Así, si un documento d recuperado por la consulta Q obtiene un valor $rel(d)$, se sustituye este valor por la media de los valores de relevancia obtenidos por aquellos documentos más similares a d en la colección, donde los documentos similares a d se obtienen utilizando distintas técnicas de agrupamiento. En base a estos valores de relevancia se construye un nuevo ranking R' y de forma equivalente a los casos anteriores se mide la diferencia entre el ranking original R y este nuevo ranking. Se

considera que la consulta es muy específica si las diferencias entre R y R' son mínimas o lo que es equivalente, que documentos similares obtengan valores de relevancia similares. Una nueva propuesta realizada por Vinay et al. (2008) es muy similar a ésta, salvo que en lugar de utilizar la media de los valores de relevancia de aquellos documentos más similares, se utiliza una distribución de probabilidad normal para calcular los nuevos valores de relevancia $rel'(d)$.

Otras aproximaciones

En Carmel et al. (2006) se plantea que la calidad de una consulta es consecuencia de los tres componentes básicos en la recuperación: consultas, documentos relevantes y la colección. Carmel define distintas medidas de distancia entre estos componentes que le permiten observar la relación existente entre las distancias y la precisión media AP que obtiene una consulta. Finalmente, la información obtenida a partir de un corpus de entrenamiento es utilizada para entrenar una máquina de vectores de soporte que haga las veces de método de predicción.

De nuevo Vinay et al. (2006) introducen una nueva técnica basada en el grado de similitud de los documentos devueltos en una consulta. Así, si al aplicar un método de agrupamiento automático al conjunto de documentos recuperados estos se agrupan de forma homogénea y robusta, se infiere una alta calidad en la lista de resultados. Por el contrario, la falta de agrupamiento entre los documentos implica cierto grado de "aleatoriedad" en la recuperación y por tanto que la consulta no será respondida con suficiente grado de calidad.

Finalmente, Aslam y Pavlu (2007) proponen comparar las listas ordenadas de documentos que se obtienen como resultado de lanzar una consulta a distintos motores de búsqueda. Así, si el orden de los documentos es similar para un gran número de los motores de búsqueda empleados, se infiere que la consulta se podía responder de forma fácil. Por contra, al observar órdenes distintos se concluye que cada motor de búsqueda ha actuado de forma muy diferente para una misma consulta, lo que indica la complejidad a la hora de resolver dicha consulta. Para medir las diferencias entre los órdenes propuestos por cada motor de búsqueda, los autores sugieren el uso de la divergencia *Jensen-Shannon*.

Extensiones a *Clarity Score*

Recientemente y en relación con el método considerado de referencia *Clarity Score*, Hauff en sus trabajos Hauff et al. (2008b); Hauff (2010) presenta dos extensiones que mejoran significativamente el rendimiento de dicho método. De hecho, el grado de correlación que se obtiene con la inclusión de dichas extensiones es uno de los más altos que ha aparecido en la literatura

relacionada.

Las limitaciones del método original, que los autores describen en estos trabajos, son aquellas relativas a la sensibilidad que *Clarity Score* presenta en relación a los parámetros que utiliza. Este hecho tiene como consecuencia, que para valores distintos de parámetros aparezcan diferencias significativas en cuanto al acierto en la predicción. Los parámetros que muestran este comportamiento son aquellos relativos a:

- El número de documentos considerados relevantes (R) a partir de la consulta original Q, con la que se estima el modelo de lenguaje de la consulta.
- El número de términos que se extraen de R para estimar el mismo modelo de lenguaje.

Para resolver ambos problemas los autores proponen dos soluciones. En primer lugar sugieren usar solo aquellos documentos en los que aparezcan todos los términos de la consulta para estimar el modelo de lenguaje de la consulta. De esta forma, se pretende eliminar del conjunto de documentos aquellos que pudieran introducir ruido y por tanto perjudicar la estimación del modelo de lenguaje de Q. En aquellos casos en los que esta restricción provoque que el número de documentos recuperados sea inferior a un límite, los autores proponen utilizar aquel conjunto de documentos de menor tamaño que se obtenga de las distintas combinaciones posibles a partir de los términos originales.

En segundo lugar, los autores proponen estimar el número adecuado de términos a utilizar en la construcción del modelo de lenguaje mediante la realización de un estudio estadístico que permita establecer dicho valor. Este estudio mide las diferencias en rendimiento cuando son usados aquellos términos contenidos en R y que a su vez aparecen en el 1, 10 o 100 % de los documentos de la colección. Se observa un mejor rendimiento con la utilización de aquellos términos que aparecen en un menor porcentaje de documentos, lo que está relacionado con el grado de especificidad de los términos, es decir aquellos términos que aparecen sólo en el 1 % de la colección.

Además de los métodos descritos en este apartado existen otras aproximaciones que no se han reflejado en este trabajo. Para realizar un recorrido más exhaustivo sobre un conjunto más amplio de métodos de predicción, se puede consultar la tesis presentada por Claudia Hauff (Hauff (2010), pág. 21-73). También en Carmel y Yom-Tov (2010) pág. 15-37, se puede encontrar un análisis de algunos de los métodos de predicción más destacados en la actualidad.

Después de haber realizado un repaso por las distintas aproximaciones de predicción más destacadas, la siguiente sección está dedicada al análisis

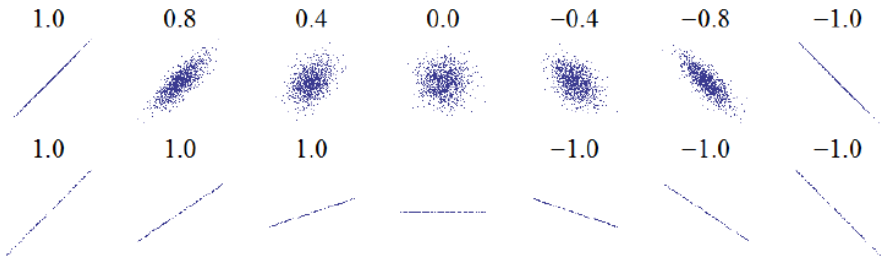

Figura 2.4: Grados de correlación para distintas series de datos (Imagen tomada de *Wikimedia Commons*: `http://en.wikipedia.org/wiki/File:Correlation_examples.png`).

del marco de evaluación que se aplica en este área para evaluar de manera cuantitativa la precisión de las predicciones.

2.4.3. Marco de evaluación de los métodos de predicción

Como se ha descrito previamente, el grado de acierto de un método de predicción se evalúa en base a la correlación encontrada, entre las estimaciones y los valores de calidad obtenidos a partir de alguna medida de evaluación como *AP*. Una primera consecuencia de este tipo de evaluación es la imposibilidad de evaluar la calidad de la predicción sobre una única consulta, ya que es necesario especificar un conjunto de consultas para poder aplicar distintos coeficientes de correlación.

Los distintos coeficientes de correlación se aplican sobre dos variables aleatorias y miden la relación, y en algunos casos dependencia, entre ambas variables. La relación entre ambas variables se cuantifica, en general, con un número real en el rango $[-1, 1]$, donde 1 implica una correlación perfecta de signo positivo y -1 se interpreta como una correlación perfecta pero de carácter inverso. Así, para valores de correlación cercanos a cero se puede concluir que no se observa dependencia entre ambas variables, aunque este hecho no implica que ambas variables aleatorias sean independientes. Por otro lado, si se observan valores de correlación altos, positivos o negativos, sugiere, aunque no asegura, una posible dependencia entre ambas variables. En la Figura 2.4, se pueden observar algunos ejemplos del grado de correlación *Pearson* que obtienen distintas series de datos dibujadas sobre un plano.

En el contexto de la predicción de la calidad de las consultas, los coeficientes de correlación que se han usado con mayor frecuencia corresponden

a los métodos de *Pearson* y *Kendall* [4]. Es importante destacar que aunque ambos métodos evalúan el grado de relación entre ambas series de datos, lo realizan de modo completamente diferente y por tanto el coeficiente de correlación que se obtiene para series de datos equivalentes con ambos métodos, puede diferir de forma significativa. Por esta razón, los resultados que se obtienen con su aplicación no son comparables. A continuación se describen en detalle las distintas características de ambos coeficientes de correlación, analizando como afecta su uso a la evaluación en el marco de la predicción del rendimiento de consultas.

Pearson

Es un método paramétrico, que asume la existencia de una relación lineal entre ambas series de datos y basándose en esta suposición mide el grado y dirección de esta relación. En otras palabras, se puede decir que el objetivo de *Pearson* es cuantificar si ambas series de datos son generadas por dos funciones que sean linealmente dependientes y en qué grado. El valor de correlación obtenido con *Pearson* se denota como r en el caso de ser calculado sobre una muestra.La ecuación más habitual que define el cálculo de r para dos series de datos $X = x_1, x_2, \ldots, x_n$ e $Y = y_1, y_2, \ldots, y_n$ se define como a continuación:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{Y})^2}}$$

Una de las interpretaciones clásicas que se da al valor que se obtiene al calcular el coeficiente de *Pearson*, corresponde con la media de las variaciones de la distancia de cada uno de los puntos en el plano a aquella línea recta que minimiza dicha distancia, siendo dicha recta la denominada recta de regresión (Rodgers y Nicewander (1988)).

Es bien conocido el hecho de que el coeficiente de correlación *Pearson* es sensible a ciertas características de los datos que pueden llevar a una interpretación equivocada de los resultados obtenidos (Devlin et al. (1975)). Así la presencia de valores atípicos en la serie de datos, es decir aquellos que se alejan considerablemente de la media, desvía de forma dramática el grado de correlación que se obtiene. Por otro lado, *Pearson* está enfocado al caso de relaciones lineales por lo que no es capaz de capturar de forma correcta los casos en los que la relación entre ambas series de datos sea no lineal.

Estas características inherentes al propio método conducen, en algunos casos, a la obtención de resultados que no reflejan la realidad de los datos. Este hecho aparece claramente recogido en el denominado *Cuarteto de Anscombe* (Anscombe (1973)). El *Cuarteto de Anscombe* consiste en un conjunto

[4] Inicialmente algunos trabajos presentaron sus resultados en base al coeficiente de *Spearman* aunque su uso actualmente se ha reducido en gran medida.

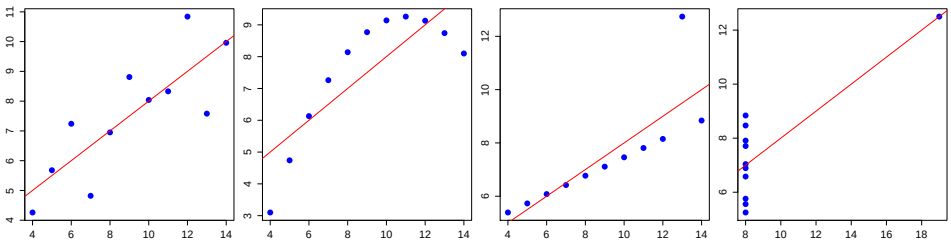

Figura 2.5: Cuarteto de Anscombe.

de cuatro pares de series de datos, donde cada una de las series obtiene el mismo valor para ciertos estadísticos como la media, varianza y especialmente el coeficiente de correlación r entre los pares. Este hecho podría llevarnos a concluir la existencia de similitudes de cierta importancia entre los pares propuestos. Sin embargo, una observación visual de los datos, como aparece en la Figura 2.5, permite apreciar que las características y naturaleza de los datos difieren en gran medida y que el grado y dirección de la relación en algunos casos es mínima. Las características estadísticas del *Cuarteto de Anscombe*, se han convertido en el ejemplo más típico para demostrar de forma empírica las limitaciones de *Pearson*, cuando es aplicado a algunas series de datos. Esto implica que los resultados obtenidos con la aplicación de *Pearson* como coeficiente de correlación, deban ser tomados con ciertas reservas a no ser que se haya realizado un profundo estudio previo sobre los datos a analizar.

Kendall

Se denota con la letra τ y su principal diferencia respecto a *Pearson* es que es un método no paramétrico. Esta condición inherente a *Kendall* implica la ausencia total de suposiciones sobre los datos, como por ejemplo que posean una relación lineal, y por tanto no requiere características específicas de estos a la hora de medir la correlación (Kendall (1938)).

El hecho de no poseer requisitos especiales sobre los datos a analizar, es la principal ventaja de *Kendall* respecto a *Pearson* en el marco de la predicción de la calidad de consultas. Por otro lado, este hecho tiene como contrapartida que *Kendall* es una medida menos informativa ya que se basa solamente en el orden relativo de los datos, obviando el valor específico de cada uno de ellos.

Kendall calcula el valor de correlación en base a los diferentes intercambios de pares de objetos realizados sobre ambas series de datos, previamente ordenadas, que serían necesarios para que las dos series tuvieran el mismo orden relativo de sus elementos.

Por ejemplo, en el caso específico de la predicción del rendimiento de

consultas, *Kendall* mide como de similares son los dos rankings resultantes de ordenar las consultas por su valor de calidad (AP) y por su valor de predicción. De esta forma, si ambos rankings están ordenados de forma equivalente se obtiene un coeficiente igual a 1. Por el contrario, si estuvieran ordenados de forma inversa uno respecto al otro, el coeficiente τ que se obtiene sería igual a -1.

Formalmente, dadas dos series de datos $X = \{x_1, x_2, \cdots, x_n\}$ e $Y = \{y_1, y_2, \cdots, y_n\}$ se dice que dos pares (x_i, y_i) y (x_j, y_j) son:

- Concordantes si:

$$x_i < x_j \text{ e } y_i < y_j \text{ o al contrario } x_i > x_j \text{ e } y_i > y_j$$

- Discordantes si:

$$(x_i - x_j)(y_i - y_j) < 0$$

- Empate si:

$$x_i = x_j \text{ o } y_i = y_j$$

Puesto que en una serie con n elementos existen $\binom{n}{2}$ pares posibles, el coeficiente τ en ausencia de empates se define como:

$$\tau = \frac{c - d}{\binom{n}{2}} = \frac{2(c - d)}{n(n-1)}$$

donde c es el número de pares concordantes, d es el número de pares discordantes y n es el tamaño de ambas muestras. Por lo que en el caso de que solo existan pares concordantes $\tau = 1$, mientras que en el caso de que sólo existan pares discordantes $\tau = -1$, es decir una correlación inversa perfecta. Finalmente $\tau = 0$ en el caso de que el número de pares discordantes sea igual al número de discordantes.

Para el caso en el que se observen empates la expresión es similar, pero se debe restar del número de pares posibles $\binom{n}{2}$, el número total de empates. Así, se define el número de empates posibles para cada uno de los rankings X e Y a partir de las expresiones que aparecen a continuación.

El número de empates posibles para el ranking X se define como T, donde:

$$T = \sum_{i=1}^{t} \frac{t_i(t_i - 1)}{2}$$

siendo t las posiciones del ranking X donde exista un empate y t_i el número de empates totales para esa posición.

De forma similar se define U como el número de empates para el ranking Y. Así el número de empates U en Y es:

$$U = \sum_{i=1}^{u} \frac{u_i(u_i - 1)}{2}$$

siendo u las posiciones del ranking Y donde exista un empate y u_i el número de empates totales para esa posición.

Finalmente, la expresión final del cómputo de τ en el caso que existan empates se define como aparece a continuación:

$$\tau = \frac{c - d}{\sqrt{\left(\frac{n(n-1)}{2} - T\right)\left(\frac{n(n-1)}{2} - U\right)}}$$

Alejándonos de los detalles más formales de la especificación de ambos coeficientes de correlación, es importante centrarse en el hecho de qué método es más adecuado y bajo que condicionantes, para el caso de la evaluación de métodos de predicción. Probablemente la característica más distintiva entre ambos métodos, es la que se refiere a qué observa cada uno de los métodos para calcular el grado de correlación. Así *Pearson* observa el grado de cercanía de los valores de AP respecto a las predicciones. Esto implica que por tanto *Pearson*, realmente se centra en el acierto a la hora de estimar el valor exacto de AP. Por otro lado, *Kendall* utiliza el orden relativo de los elementos ignorando los valores exactos que ocurren, por lo tanto mide la similitud entre ambos rankings ordenados.

Aunque idealmente se podría esperar que un método de predicción *perfecto* fuera capaz de predecir el valor exacto obtenido con la medida de evaluación objetivo, siendo este ideal el que se evalúa con la aplicación de *Pearson*, parece claro que con las técnicas actuales aún se está lejos de este ideal. Sin embargo, la evaluación de los métodos de predicción usando *Kendall* no sólo evita algunos de los problemas descritos respecto a la aplicación de *Pearson*, sino que parece un enfoque más realista y más acorde con la capacidad actual de los métodos propuestos. Como ya ha quedado reflejado, *Kendall* ignora los valores de la función objetivo centrándose sólo en el orden, por lo tanto aunque no nos permite estimar el valor exacto de calidad de una consulta, sí que muestra el grado de acierto de un método de predicción en cuanto a qué consultas son respondidas de forma más adecuada y cuales de ellas reciben una respuesta poco relevante.

Estas razones han hecho que en los últimos tiempos, la evaluación basada en *Kendall* se vaya imponiendo y que por tanto los resultados de los distintos métodos de predicción tiendan a presentarse en base a este coeficiente, o al menos utilizando ambos métodos.

Posibles extensiones al uso de coeficientes de correlación

Algunas de las características que hacen que *Pearson*, en algunos casos, no sea la medida más adecuada en el contexto de la predicción de la calidad,

fueron descritas por Hauff et al. (2009) previamente. En este mismo trabajo se proponía el uso del Error Cuadrático Medio *(RMSE)* entre ambas series, para evaluar la calidad de la predicción. El Error Cuadrático Medio es la distancia promedio entre la medida de calidad y las estimaciones, y se define como aparece a continuación:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}$$

Con esta medida se corrige el problema de la no-linealidad, aunque no resuelve el problema de la presencia de datos atípicos. Además, al ser una medida que utiliza la distancia absoluta entre ambos valores, no puede ser aplicada para comparar la efectividad de distintos métodos de predicción, algo que no ocurre con *Pearson* ya que éste no es sensible a transformaciones afines.

Por otro lado, con la aplicación de *Kendall* también existen ciertos problemas que son inherentes al mismo hecho de comparar rankings. La principal crítica se basa en el hecho de que en *Kendall* todas las posiciones del ranking se consideran de igual importancia, algo que en realidad en el contexto que nos ocupa no es lo más acertado. Así, en un ranking es muy probable que nuestro interés se centre en aquellos elementos que aparezcan en las primeras posiciones. Por ejemplo, es posible que en el contexto de la predicción de la calidad de las consultas, se tenga un mayor interés en detectar aquellas consultas que muestren un rendimiento elevado y que por tanto no sean candidatas para la aplicación de técnicas que mejoren calidad. Por contra, podría existir un especial interés en aquellas consultas que muestran un bajo rendimiento con el objetivo de mejorar su respuesta. Por tanto, parece claro que no todas las posiciones del ranking deben tener la misma influencia en la medida final.

Esta razón, en el contexto de la comparación de rankings, ha llevado a diversos autores a proponer aproximaciones que permitan definir distintos grados de importancia de los elementos, según la posición que ocupan en el ranking. Estas extensiones se denominan de forma genérica *Weighted Kendall*.

El grado de importancia que se asigna a cada uno de los elementos se puede asignar de forma explícita, por ejemplo en base a un factor constante k que incrementa la influencia de ciertos elementos en nuestro ranking en la medida final de correlación que se obtenga, como se propone en Melucci (2007). Otra aproximación consistiría en asignar la importancia relativa de cada elemento en el ranking según una función específica para este cometido, como sugieren los trabajos realizados por Yilmaz et al. (2008) o Melucci (2009). Aunque se ha mostrado interés en este tipo de medidas y en su adecuación al problema de la predicción (Hauff (2010), pág. 148,149.), en la

actualidad no existen trabajos que hayan publicado sus resultados en base a estas medidas.

Es destacable que la evaluación de los distintos métodos de predicción es un campo abierto que no ha sido tratado en profundidad, aunque existe un claro acuerdo en la comunidad respecto a las limitaciones que presenta la evaluación basada en coeficientes de correlación (Hauff (2010), pág. 148,149), (Carmel y Yom-Tov (2010), pág.13).

2.5. Conclusiones

Finalmente y como conclusión al amplio recorrido desarrollado sobre las distintas tecnologías que intervienen en el trabajo propuesto, se especifican algunas de las principales conclusiones a las que se ha llegado y que pueden servir como síntesis del estado actual de este campo de investigación, que se resumen en los siguientes puntos:

- Los métodos *Pre-Retrieval* obtienen valores de correlación inferiores a aquellos obtenidos mediante la utilización de métodos *Post-Retrieval* (Hauff (2010), cap. 2).
- Los métodos *Pre-Retrieval* que obtienen mejores resultados son aquellos que se basan en la especificidad de los términos de la consulta, en base a estadísticos como *IDF* o *ICTF*. En general con ambas aproximaciones se obtienen resultados muy similares He y Ounis (2004, 2006).
- Algunos de los métodos *Post-Retrieval* poseen una alta complejidad, lo que en muchos de los casos imposibilita su aplicación en entornos reales.
- En general, existe un gran similitud entre los resultados obtenidos con los distintos métodos de predicción en términos de correlación. Este hecho dificulta en gran medida una correcta comprensión de las diferencias de capacidad de predicción de cada uno de ellos. Por tanto, sería deseable la existencia de otros métodos de evaluación que facilitaran la observación de las diferencias de rendimiento entre distintas técnicas de predicción, y así poder realizar una selección más adecuada de estos métodos dependiendo del ámbito donde se deseen aplicar.
- Los resultados mostrados en los distintos artículos son muy dependientes de la colección donde se realice y del método de correlación seleccionado para su evaluación.

A fact is a simple statement that everyone believes. It is innocent, unless found guilty. A hypothesis is a novel suggestion that no one wants to believe. It is guilty, until found effective.

Edward Teller.

Capítulo 3

Predicción sobre rankings de documentos

3.1. Introducción

En este capítulo se describe una nueva aproximación a la predicción del rendimiento de consultas. Esta aproximación se basa únicamente en los distintos pesos calculados por una función de ranking. Dichos pesos o valores de relevancia corresponden a aquellos que se asignan a cada uno de los documentos devueltos por un sistema de información al responder a una consulta. En la literatura relacionada se pueden encontrar diversos métodos que utilizan los pesos asignados por la función de ranking como estimadores. Así dentro de este grupo destacan:

- Diaz (2007), donde a partir del ranking R obtenido dada una consulta, se construye otro ranking R' con los mismos documentos pero modificando los valores de relevancia que aparecían en R. Los nuevos valores de relevancia de R' se calculan basándose en los valores que tenían los documentos según R en combinación con los valores de relevancia asignados a otros documentos similares. Por tanto esta técnica implica ejecutar algún algoritmo de agrupamiento, tal que dado un documento $d \in R$ se puedan obtener documentos similares dentro de la colección.

- Vinay et al. (2008), donde se propone un enfoque muy similar al descrito en el punto anterior. La diferencia principal reside en el hecho de que los nuevos valores de relevancia en R', se generan mediante la estimación de una distribución normal de los pesos. Por tanto, esta técnica añade la dificultad de estimar una distribución de probabilidad que se ajuste a los datos originales.

- Zhao et al. (2008), que proponen un método de tipo *Pre-Retrieval*. Sin embargo, es tan costoso como los típicos métodos *Post-Retrieval* ya que

debe calcular un valor basado en un esquema de pesado *TF-IDF*, por cada documento de la colección en el que aparezca al menos uno de los términos de la consulta. Esta tarea supone un coste similar a recuperar los documentos dada una consulta, en un sistema de recuperación que incluya como modelo de relevancia un esquema *TF-IDF*.

La utilidad del método presentado tiene dos características que combinadas lo hacen relevante. Por un lado, muestra un rendimiento similar, en un gran número de casos, a otras aproximaciones que son consideradas en la actualidad como las más efectivas, tales como *Clarity Score* o los métodos mencionados previamente. De forma simultanea, basa sus estimaciones en la aplicación de operaciones muy sencillas lo que permite su ejecución con un bajo coste computacional.

El resto del capítulo se estructura de la siguiente forma: en primer lugar se describen las distintas hipótesis que dan fundamento al método de predicción introducido. Posteriormente, se realiza un análisis en detalle de los dos factores que pueden influir en el rendimiento de la predicción, siendo estos la normalización de los valores de relevancia y la selección del conjunto de documentos sobre los que se calcula la dispersión entre los valores de relevancia. Dicha selección se realiza estableciendo un punto de corte en la lista de resultados que se obtiene a partir de una consulta.

Se presentan diversos métodos para realizar la selección de dicho punto de corte, para finalizar con los resultados que se obtienen con esta nueva aproximación. Finalmente, se da explicación al bajo rendimiento general de los distintos métodos de predicción sobre la colección *WT10g*.

Las funciones de ranking actuales, en contraposición a aquellas como la mera recuperación booleana, se sirven de un conjunto de estadísticos, a nivel de documento o de colección, para asignar un valor de relevancia a cada uno de los documentos recuperados. Dichos estadísticos se obtienen en base a los términos de la consulta del usuario. Así, entre los estadísticos más típicos que se utilizan, podemos considerar la frecuencia de los términos de la consulta en la colección, la frecuencia de dichos términos en cada uno de los documentos candidatos a ser recuperados, la longitud de los documentos o la longitud de la colección. Como se ha mencionado a lo largo de esta tesis, el conjunto de documentos devueltos a partir de una consulta serán ordenados de forma decreciente según el valor de relevancia que obtenga cada documento. Por tanto, esta lista ordenada será la que se devuelva al usuario como respuesta a una consulta específica. Dependiendo del número de documentos relevantes para el usuario que contenga dicha lista, así como la posición de estos en el ranking, se considerará la respuesta como más o menos acertada.

La principal limitación inherente a este tipo de aproximaciones es que un documento es recuperado dependiendo de la presencia de los términos de la consulta en dicho documento. Sin embargo, la mera presencia de los términos

de una consulta en un documento puede ser un indicativo de relevancia pero no asegura la relevancia de dicho documento. Así por ejemplo, las funciones de ranking tienden a asignar un valor igual o cercano a cero a aquellos documentos en los que no aparezca ninguno de los términos de la consulta original. Como consecuencia, estos documentos suelen aparecer muy alejados de las primeras posiciones en el ranking sean o no relevantes para el usuario.

Por tanto podemos ver a una función de ranking como una función cuyo objetivo es separar documentos relevantes de aquellos no relevantes. Para realizar esta tarea la función de ranking debe asignar pesos diferentes de forma que los menos relevantes aparezcan al final del ranking y los relevantes al principio. Así, si una función de ranking se demuestra muy efectiva para una consulta específica, es de esperar que los valores más altos correspondan a documentos relevantes mientras que los valores más bajos aparezcan en los documentos no relevantes, estableciendo una clara diferencia entre dichos valores.

Aumentar las diferencias entre los distintos valores de relevancia asignados por una función de ranking, tiene como consecuencia un importante incremento en la dispersión de los valores de relevancia. Por tanto, el grado de dispersión entre los valores de relevancia puede ser interpretado como un criterio de la calidad de la respuesta a una consulta dada. En el caso que se observe un alto grado de dispersión podemos suponer que la consulta ha obtenido una respuesta de calidad, ya que la función de ranking ha separado los documentos relevantes de los no relevantes. Sin embargo, si la dispersión es baja puede ser consecuencia de que la función de ranking considera a todos los documentos devueltos equiprobables en términos de relevancia respecto a la consulta.

Esta aproximación a la predicción de la calidad de las consultas viene limitada por la propia naturaleza de las funciones de ranking. Como se ha mencionado con anterioridad las diferencias entre los estadísticos no siempre describen exactamente el grado de relevancia de un documento, y como consecuencia las dispersiones que aparezcan entre los documentos no reflejarán las diferencias reales en términos de relevancia. Este hecho provoca que la dispersión como criterio de la calidad de una respuesta depende en gran medida del rendimiento mostrado por la función de ranking.

Dispersión entre los valores de relevancia

La dispersión sobre una muestra de datos puede ser medida a través de la desviación estándar σ, que mide la variabilidad de un conjunto de datos. Un valor bajo de dispersión indica que los datos medidos tienden a estar muy cercanos a la media μ. Simultáneamente, un valor alto de desviación implica que el conjunto de valores de estudio se encuentra disperso sobre un amplio rango de valores distintos. Dada una lista de documentos ordenada por sus pesos, la desviación estándar, para el caso poblacional, calcula el

promedio de las distancias de cada uno de los puntos a la media aritmética de dichos puntos, según la siguiente ecuación:

$$\sigma(RL) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(score(d_i) - \mu(RL))^2}$$

donde $\mu(RL)$ corresponde con la media de los valores de relevancia y N es el tamaño de la lista de documentos.

El significado de la dispersión entre los valores de relevancia se pone de manifiesto más claramente si se considera un modelo teórico de recuperación perfecto. En esta aproximación teórica se supone que la función de ranking posee una efectividad completa. Es decir, siempre asigna uno a los documentos que son realmente relevantes y cero a los documentos no relevantes.

En este caso teórico se maximizaría la dispersión en la lista de documentos recuperados en el caso de que el número de documentos relevantes fuera igual al número de documentos no relevantes. Por tanto, para un tamaño de ranking igual a k la dispersión entre los valores de relevancia sería máxima si:

$$numRelevantes = numNoRelevantes = \frac{k}{2}$$

Aunque ninguna función de ranking muestra un rendimiento tan óptimo como el modelo teórico descrito, algunas funciones de ranking permiten observar un mayor grado de dispersión entre los valores de relevancia asignados a consultas de mayor calidad.

Este hecho aparece de forma gráfica en la Figura 3.1, donde se muestran claras diferencias en el grado de dispersión de los valores de relevancia, asignados por una función de ranking clásica como *BM25*, entre consultas que obtienen valores muy distintos en términos de precisión media. En esta figura, se comparan las cinco consultas que obtienen un mejor rendimiento, Tabla 3.1 ($AP > 0{,}5$), con las cinco consultas con menor rendimiento, Tabla 3.2 ($AP < 0{,}05$), para la tarea *Robust2004*. Al normalizar los pesos en el rango $[0, 1]$, se pueden observar importantes diferencias entre ambos grupos de consultas. Así, aquellas consultas que obtienen un mejor resultado en términos de precisión media (AP) muestran una diferencia mayor entre los documentos del principio del ranking y aquellos que aparecen al final, lo que genera gráficas con mayor pendiente. Por el contrario, las consultas con baja precisión media muestran una mayor estabilidad en el conjunto de los valores asignados por la función de ranking. En base a la hipótesis planteada previamente, se observa que en algunos casos, la dispersión entre los valores de relevancia de las consultas con una precisión media más elevada, es superior a la dispersión que muestran los valores de relevancia de las consultas con menor precisión media.

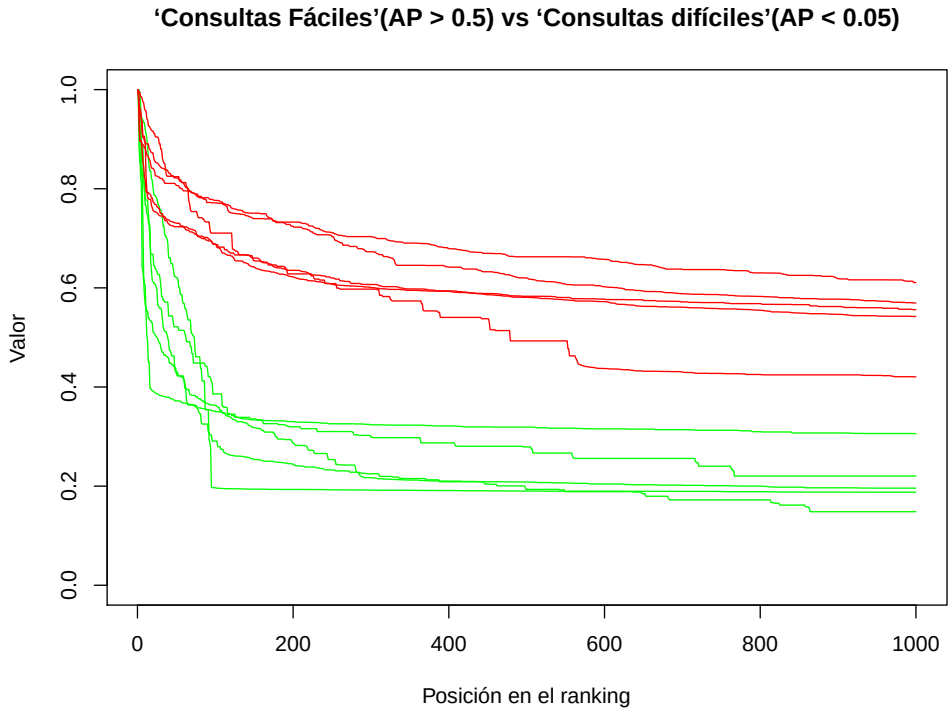

Figura 3.1: , entre las cinco consultas con mejor rendimiento (en verde) y las cinco con peor rendimiento (en rojo) en la tarea *Robust2004* (Voorhees (2004)). El valor asignado a cada documento ha sido previamente normalizado entre el rango $[0, 1]$. Para cada consulta se han tenido en cuenta los 1000 primeros documentos recuperados.

Tabla 3.1: Cinco consultas pertenecientes a la tarea *Robust2004* que obtienen mayor valor de precisión media usando *BM25* como función de ranking.

Consulta	**Title**	**AP**
313	*"Magnetic Levitation-Maglev"*	0.6645
365	*"El Nino"*	0.6603
410	*"Schengen agreement"*	0.8553
444	*"supercritical fluids"*	0.6458
604	*"Lyme disease arthritis"*	0.6819

Tabla 3.2: Cinco consultas pertenecientes a la tarea *Robust2004* que obtienen menor valor de precisión media usando *BM25* como función de ranking.

Consulta	Title	AP
305	*"Most Dangerous Vehicles"*	0.0012
309	*"Rap and Crime"*	0.0004
314	*"Marine Vegetation"*	0.0022
318	*"Best Retirement Country"*	0.0040
322	*"International Art Crime"*	0.0009

Partiendo de la hipótesis planteada junto a los datos observados, es posible definir un método de predicción, en base a la dispersión entre los valores de relevancia asignados por una función de ranking, suponiendo que se cumple lo siguiente:

- El valor asignado por una función de ranking a un documento trata de establecer el grado de relevancia de dicho documento respecto a la consulta.
- Aunque en algunos casos la respuesta obtenida a partir de una consulta puede ser completamente irrelevante para el usuario, en general y en caso de que los hubiera, los documentos relevantes se sitúan al principio del ranking devuelto.
- De la misma forma la mayoría de los documentos no relevantes se sitúan al final del ranking, ya que estos tendrán poca relación con el contenido de la consulta.
- Dado el tamaño de las colecciones de evaluación actuales, podemos suponer que el número de documentos relevantes es mucho menor que el número de documentos no relevantes.
- Existen documentos relevantes y no relevantes con valores similares y que por tanto tienen posiciones similares en el ranking.

Inicialmente existen dos factores de importancia que pueden afectar a los valores de la desviación estándar medidos sobre un ranking de documentos. Estos son, la utilización de algún método de normalización sobre los valores de relevancia y el número de documentos o tamaño de la lista de resultados sobre la cuál se calcula la desviación.

3.2. Normalización de los valores de relevancia

En diversas situaciones en el campo de la recuperación de información, es necesario comparar los valores de relevancia en un conjunto de documentos.

Así, puede ser que deseemos comparar los valores obtenidos con funciones de ranking distintas para una misma consulta, o simplemente comparar los valores que obtienen dos consultas diferentes.

En la mayoría de los casos una comparación directa de los valores de relevancia puede conducir a una interpretación errónea, ya que los valores que se obtienen dependen de factores no directamente relacionados con la relevancia, como la longitud de las consultas o las diferencias en los estadísticos de colecciones distintas. Así, aquellas consultas de mayor longitud generalmente recuperan documentos con valores de relevancia más elevados que aquellas con menor número de términos, cuando se utiliza un mismo corpus para ambas.

Este efecto es consecuencia de la propia definición de las funciones de ranking. Por ejemplo, en los casos de modelos de recuperación aplicados en esta tesis como *BM25* o *Query Likelihood*, el valor final que obtiene un documento se calcula como un sumatorio de los valores obtenidos por cada término de la consulta. Por tanto, a mayor número de términos en la consulta mayor será la probabilidad de que algún documento los contenga y de obtener un valor de relevancia superior.

Sin embargo, no se puede afirmar que las consultas más largas tiendan a recibir una respuesta más relevante, ya que un valor de relevancia más elevado no implica una relevancia mayor para el usuario.

A priori y en base a la especificidad de una consulta, es intuitivo concluir que aquellas consultas de mayor longitud obtendrán precisiones medias más altas, ya que son más descriptivas respecto al asunto del que el usuario desea recuperar documentos. Sin embargo, la aparición de muchos términos en una consulta puede provocar el efecto contrario, ya que algunos de estos pueden desviar el asunto central de la necesidad informativa del usuario. De forma similar, es esperable que aquellas consultas con un único término sean más difíciles que aquellas con mayor número de términos. De nuevo este hecho no siempre ocurre ya que una consulta de un término muy específico (*IDF* alto) será más fácil de responder que una consulta con muchos términos muy generales (*IDF* bajo).

Los distintos valores de relevancia asignados por una función de ranking afectarán al valor de dispersión que obtenga una consulta. Se puede esperar que al comparar la dispersión que muestran dos listas de documentos distintas, aquella que se haya obtenido con una consulta más larga refleje por lo general un mayor grado de dispersión. Por la misma razón y como consecuencia, aquellas consultas más largas serán normalmente predichas como consultas que obtienen una precisión media más elevada, lo que puede ser atenuado mediante el proceso de normalización de los valores de relevancia.

La utilización de valores de relevancia normalizados permitirá comparar los valores de relevancia sin primar a aquellas consultas de mayor longitud. La característica principal del método de normalización a aplicar debe ser que modifique la escala de los datos de forma equitativa y que por tanto no

introduzca distorsiones entre los valores. Así, debe mantener las distancias relativas entre valores de relevancia para una misma consulta. Cualquier modificación en estas distancias afectaría a lo establecido en primer término por la función de ranking.

Esta restricción hace que se descarten métodos de normalización clásicos como *z-score* y *min-max*, ya que ambos métodos modifican la media de los datos manteniendo constante la desviación. Esto supone que cada valor de relevancia será escalado por un valor distinto que depende de su distancia a la media. Por tanto, las distancias entre los documentos son modificadas en distinto grado, con el objetivo de mantener la desviación constante en un conjunto de datos donde la media varía.

Dado un conjunto de datos $x_1, x_2, ..., x_n \in X$, el método de normalización *z-score* se define según la siguiente ecuación:

$$z - score_i = \frac{x_i - \mu(X)}{\sigma(X)}$$

De forma similar *min-max* se define:

$$min - max_i = \frac{x_i - min(X)}{max(X) - min(X)}$$

Un método clásico de normalización que cumple con la condición expuesta, es aquel que consiste en dividir cada elemento de la serie de datos por el máximo en valor absoluto de dicha serie, tal como aparece a continuación:

$$x'_i = \frac{x_i}{|max(X)|}$$

Este método da como resultado un nuevo conjunto de datos definido en el intervalo $[\frac{min(X)}{max(X)}, 1]$, donde los nuevos valores de la media y la desviación corresponden a la media y desviación originales divididos por el máximo en valor absoluto.

Con el objetivo de comprobar el efecto que produce la normalización de los valores de relevancia a la hora de predecir la calidad de las consultas en base a la dispersión, se plantean los siguientes escenarios:

- No realizar normalización.
- Aplicar normalización por el máximo valor encontrado en cada consulta.

3.2.1. No normalización

No realizar el proceso de normalización implica mantener los valores asignados a cada uno de los documentos tal y como son calculados por la función de ranking. Por tanto, se mantiene el sesgo que relaciona la longitud

Tabla 3.3: Número de términos que contienen las consultas de las colecciones de referencia.

#términos	**TREC Vol. 4+5**	**WT10g**	**GOV2**
1	13	19	3
2	85	27	46
3	139	26	63
>3	12	25	37
Total	249	97	149

de una consulta con la calidad de ésta. La hipótesis que sustenta la omisión del proceso de normalización es que las consultas más largas obtienen una respuesta más adecuada por ser más descriptivas. Esta hipótesis se contrapone a la idea de que las consultas cortas suelen ser de un carácter más específico y por tanto recuperan documentos relevantes con mayor facilidad.

En aquel conjunto de consultas y colecciones en las que se detecte una relación directa entre la longitud de una consulta y la calidad de la respuesta que obtiene, no realizar el proceso de normalización debería mejorar el rendimiento del método de predicción.

3.2.2. Normalización por el máximo

Normalizando se intenta corregir el hecho de que aquellas consultas que por su longitud obtienen valores de relevancia más elevados, producirán a su vez dispersiones mayores. Por tanto, al omitir la normalización se evita asumir una relación directa entre la longitud de una consulta y su calidad.

Un primer paso para analizar la idoneidad de realizar el proceso de normalización, es estudiar qué tipo de relación existe entre la longitud de las consultas y el valor de precisión media que se obtiene. En el caso de que exista dicha relación, es decir que a mayor longitud de la consulta se encuentra, en general, valores de AP más altos, el proceso de normalización sería contraproducente ya que reduce la desviación de aquellas consultas más largas. Si por el contrario, nos encontramos con que aquellas consultas más cortas obtienen valores de AP más altos, la normalización sería adecuada ya que se disminuye en mayor medida el valor de desviación en aquellas consultas de mayor longitud.

En la Tabla 3.3 se puede observar el número de términos de las consultas que componen las colecciones de test usadas en esta tesis.

A su vez la Tabla 3.4 refleja los valores de MAP que se obtienen para aquellas consultas con uno, dos, tres o más términos. Los resultados parecen indicar que para el caso de las consultas empleadas en la tarea *Robust2004*, aquellas con un único término obtienen un valor de AP superior al resto de consultas que contienen más de un término. Una posible razón a este hecho es el alto grado de especificidad de algunas consultas con un solo término, tales

Tabla 3.4: utilizando *BM25* y *Query Likelihood (QL)* como funciones de ranking, donde T45 es *TREC Vol. 4+5*, W es *WT10g* y G corresponde a *GOV2*.

#términos	$\mathbf{T45_{BM25}}$	$\mathbf{T45_{QL}}$	$\mathbf{W_{BM25}}$	$\mathbf{W_{QL}}$	$\mathbf{G_{BM25}}$	$\mathbf{G_{QL}}$
1	0.3648	0.3989	0.1451	0.1549	0.1149	0.1172
2	0.2189	0.2322	0.2545	0.2676	0.2909	0.2898
3	0.2328	0.2360	0.1712	0.1928	0.3029	0.3021
>3	0.2180	0.2108	0.2222	0.2013	0.2846	0.2762

como la consulta 312, *"Hydroponics"*, o la 348, *"Agoraphobia"*. Para el caso de las consultas realizadas sobre la colección *WT10g*, no es posible extraer una conclusión clara, ya que parece que aquellas consultas con uno o tres términos obtienen valores de *AP* inferiores respecto a las consultas de dos o más de tres términos, y por tanto no aparece una tendencia reconocible. Algo similar ocurre con las consultas en *GOV2*. En este caso se observa claramente que las consultas con solo un término muestran un rendimiento muy inferior al del resto de consultas. Sin embargo, este hecho se observa sobre un número tan reducido de consultas, tan solo tres, que no permite asegurar que la aplicación del proceso de normalización beneficiaría en gran medida a la calidad de las predicciones realizadas.

El efecto que produce el proceso de normalización de cara a mejorar la calidad de las predicciones aparece de forma gráfica en las Figuras 3.2, 3.3, 3.4, 3.5, 3.6 y 3.7, donde se muestra el grado de correlación entre la desviación estándar y el valor de *AP* que obtienen las consultas del conjunto de colecciones. El análisis se realiza utilizando las funciones de ranking *BM25* y *Query Likelihood* y evaluando respecto a dos coeficientes de correlación *Pearson* y *Kendall*.

En el caso de las consultas de la tarea *Robust2004* (Figuras 3.2 y 3.3), se observa una mínima mejoría si se aplica el proceso de normalización al evaluar con *Pearson*. Sin embargo, no se observa una tendencia clara al aplicar el coeficiente *Kendall*, lo que impide extraer conclusiones respecto a la conveniencia de aplicar el proceso de normalización a los valores de relevancia.

Algo similar ocurre al observar los datos de correlación en la colección *WT10g*, (Figuras 3.4 y 3.5). De nuevo las diferencias que aparecen son mínimas tanto en el caso de usar la correlación *Pearson* como *Kendall*. Estas pequeñas diferencias sugieren cierta mejora en la calidad de las predicciones si no se realiza el proceso de normalización.

Finalmente y al analizar el rendimiento mostrado en la colección *GOV2*, en las Figuras 3.6 y 3.7, aparecen diferencias poco significativas al aplicar o no el proceso de normalización. Este hecho, puede deberse a que no se detectó una tendencia clara que relacione la longitud de las consultas y la

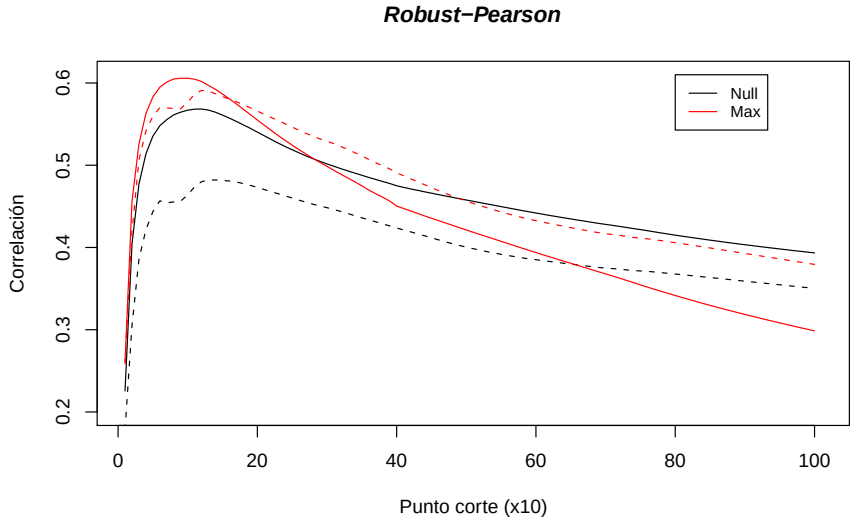

Figura 3.2: Los métodos de normalización utilizados son: No Normalización (*Null*) y Normalización por el Máximo (*Max*). En línea continúa aparece la correlación obtenida utilizando *BM25* como función de ranking y en discontinua *Query Likelihood*.

calidad de la respuesta que obtienen.

Es difícil extraer conclusiones definitivas en base a estos resultados, ya que la aplicación del proceso de normalización no produce apenas efecto en la calidad de las predicciones. Sin embargo, sí se observa en las figuras una importante influencia del número de documentos que se utilizan para calcular la desviación de los datos. Este efecto, junto con la aplicación de técnicas para estimar el tamaño de la lista de documentos óptimo que maximice el grado de correlación, se trata en detalle en la Sección 3.3.

3.3. Efecto de la longitud de la lista de documentos

A la hora de aplicar una medida de dispersión, como la desviación estándar, es importante analizar el efecto que provoca en el valor de desviación que se obtiene sobre los datos, el tipo de distribución de probabilidad que genera los valores. Obviamente la medida de dispersión se verá afectada por el tipo de distribución de los datos.

Dentro del campo de la RI se encuentran un número relevante de publicaciones, cuyo principal objetivo es estimar qué tipo de distribuciones de probabilidad son las que dan lugar a los valores de relevancia generados por

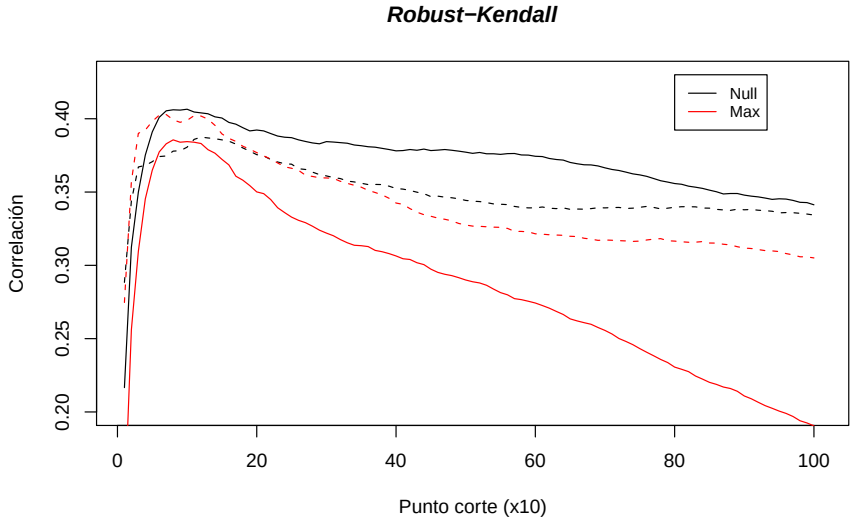

Figura 3.3: Los métodos de normalización utilizados son: No Normalización (*Null*) y Normalización por el Máximo (*Max*). En línea continúa aparece la correlación obtenida utilizando *BM25* como función de ranking y en discontinua *Query Likelihood*.

una función de ranking. Aunque no existe un consenso claro sobre el tipo o tipos de distribución de probabilidad que caracterizan una función de ranking, se considera como la aproximación más apropiada la que supone que la distribución de probabilidad viene definida mediante la composición de una distribución exponencial y una normal. Así, la distribución exponencial genera los pesos de aquellos documentos considerados no relevantes mientras que la normal se aplica a los documentos relevantes (Manmatha et al. (2001); Arampatzis y van Hameran (2001); Robertson (2007)).

La consecuencia directa de estas aproximaciones es la aparición de la denominada 'cola de la lista' que coincide con la larga cola que aparece en una distribución de probabilidad exponencial típica. Esta larga cola es consecuencia del gran número de documentos no relevantes recuperados a partir de una consulta, según las aproximaciones anteriores, y por tanto es de esperar que en dicha distribución la gran mayoría de los documentos obtengan pesos o valores bajos en comparación a aquellos generados por la distribución normal.

La aparición de dicha cola de documentos se puede observar en los histogramas de los valores de relevancia que aparecen en la Figura 3.8. Ambos histogramas corresponden a dos consultas pertenecientes a la tarea *Robust2004*, donde dichos valores han sido normalizados usando el máximo

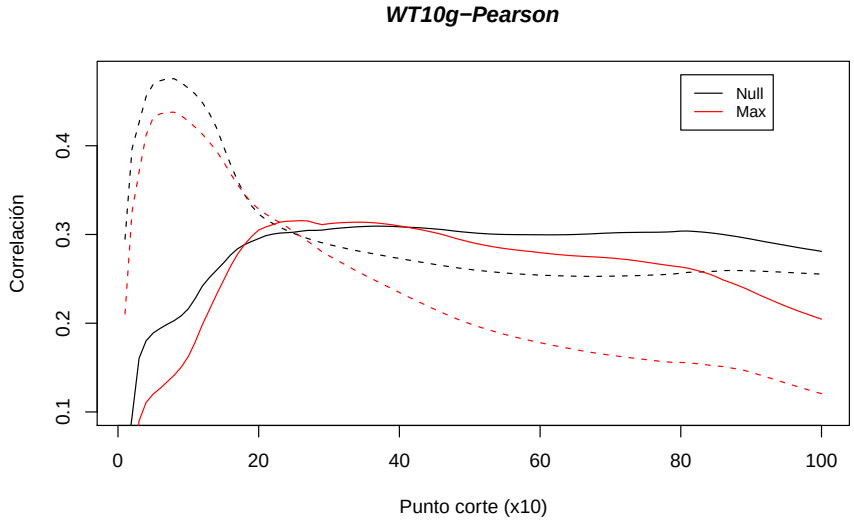

Figura 3.4: Los métodos de normalización utilizados son: No Normalización (*Null*) y Normalización por el Máximo (*Max*). En línea continúa aparece la correlación obtenida utilizando *BM25* como función de ranking y en discontinua *Query Likelihood*.

valor de relevancia. La primera consulta corresponde a la 313, *"Magnetic Levitation-Maglev"*, que puede ser considerada como de alta calidad ya que obtiene un valor muy elevado de precisión media, en el entorno del 0,7. También se muestra la consulta 305, *"Most Dangerous Vehicles"*, que muestra un bajo rendimiento, ya que obtiene una precisión media por debajo de 0,01.

Aunque las diferencias entre ambas consultas en términos de precisión media son notables, ambos histogramas muestran cierta similitud en lo que a la larga cola se refiere. Por ejemplo, en el caso de la consulta con AP alto (313), se observa que la función de ranking asigna a más de un 90 % de los documentos un valor de relevancia normalizado inferior a 0,3. De la misma manera, en relación a la consulta 305 más de un 70 % de los documentos obtienen valores de relevancia que no superan 0,7, siendo el mínimo valor de relevancia para dicha consulta igual a 0,56. Es este conjunto de documentos con valores bajos y muy similares, los que conforman la denominada 'larga cola' y que en general contiene un gran número de documentos no relevantes.

La consecuencia de la aparición de la 'larga cola' es que el valor de desviación estándar que se obtiene sobre el conjunto de los valores de relevancia esté distorsionado, debido al gran número de documentos no relevantes. Este conjunto de documentos, con valores muy bajos de relevancia, hace que la media de la muestra disminuya. El desplazamiento de la media hacia la zona

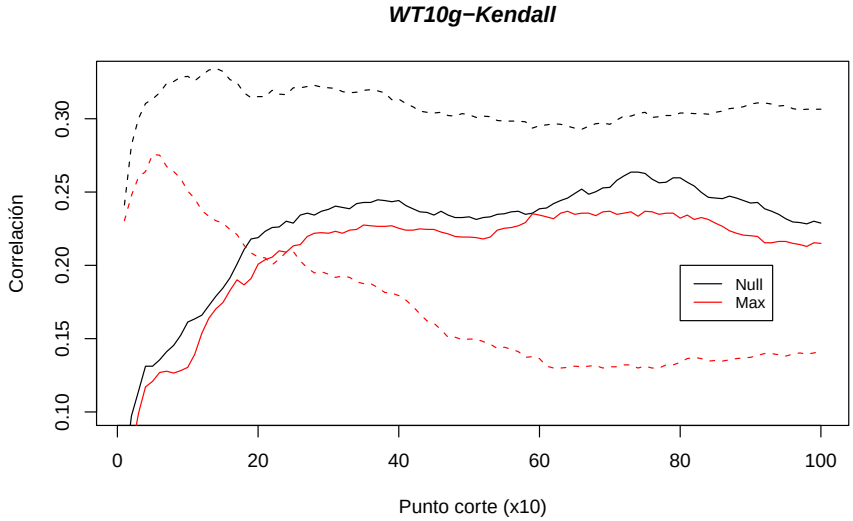

Figura 3.5: Los métodos de normalización utilizados son: No Normalización (*Null*) y Normalización por el Máximo (*Max*). En línea continúa aparece la correlación obtenida utilizando *BM25* como función de ranking y en discontinua los resultados con *Query Likelihood*.

de valores de relevancia más bajos, hace que disminuya la distancia de un gran número de valores de relevancia respecto a ésta y como consecuencia reduce el valor de desviación estándar sobre la muestra.

Este efecto se reduce si se mide la dispersión evitando introducir la cola de la lista generada por una teórica distribución exponencial, es decir, midiendo la dispersión entre aquellos k documentos que aparezcan primeros en el ranking.

Con el objetivo de seleccionar los primeros k documentos más adecuados a la hora de medir la dispersión, en la siguiente sección se proponen una serie de técnicas para calcular k de forma óptima. De esta forma, se maximiza, en lo posible, la correlación entre la desviación y la precisión media que obtiene un conjunto de documentos.

3.3.1. Medidas propuestas para el cálculo de un punto de corte común

En base a las características propias de los valores de relevancia descritos previamente, se opta por basar las predicciones en la medida de la desviación estándar pero estableciendo ciertos mecanismos que mejoren la calidad de las predicciones. En relación a la normalización, en base al estudio realizado, se opta por aplicar una normalización basada en el máximo para el

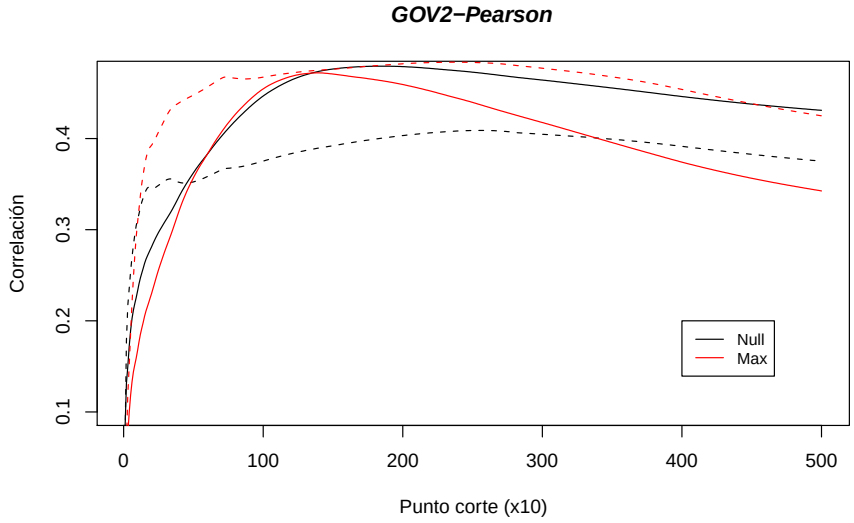

Figura 3.6: Los métodos de normalización utilizados son: No Normalización (*Null*) y Normalización por el Máximo(*Max*) . En línea continúa aparece la correlación obtenida utilizando *BM25* como función de ranking y en discontinua *Query Likelihood*.

conjunto de los valores de relevancia, aunque como se detalló previamente solo aparece un efecto levemente positivo, para el caso de las consultas de la tarea *Robust2004*.

Con el objetivo de disminuir el efecto de la cola de documentos no relevantes, se incluyen ciertos mecanismos para estimar el tamaño óptimo del ranking a partir del cual calcular la desviación.

La primera aproximación, y la más simple, mide la desviación estándar de los valores de relevancia para el tamaño de lista k, tal que se maximice el grado de correlación entre la desviación y el valor de calidad objetivo, que en general es la precisión media obtenida por cada una de las consultas.

Se utiliza la siguiente notación: la lista de documentos RL recuperados a partir de una consulta, ordenada según los valores de relevancia de forma decreciente. Además se define el valor de relevancia asignado al documento i-ésimo (d_i) en la lista de documentos como $valor(d_i)$. Por tanto dado aquel k que maximice la correlación tendríamos la siguiente ecuación

$$\sigma(RL_k) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(valor(d_i) - \mu(RL))^2}$$

Para comprobar la validez de esta primera aproximación se muestran los

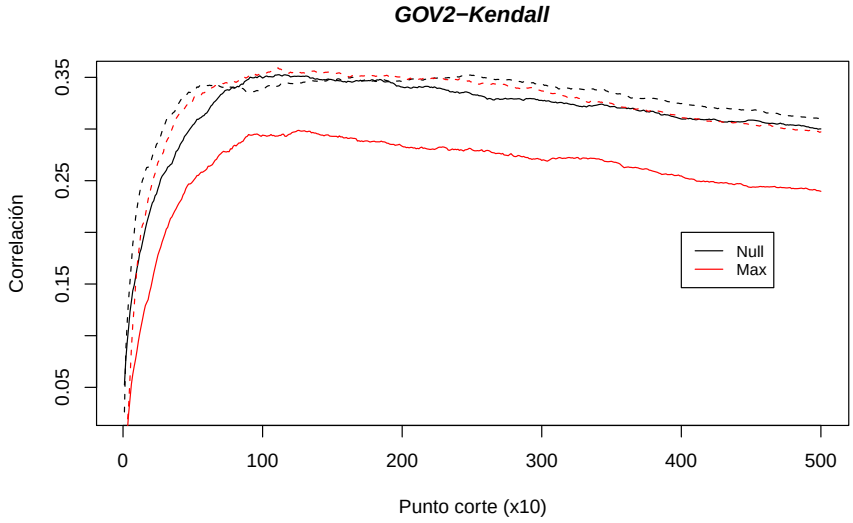

Figura 3.7: Los métodos de normalización utilizados son: No Normalización (*Null*) y Normalización por el Máximo(*Max*) . En línea continúa aparece la correlación obtenida utilizando *BM25* como función de ranking y en discontinua *Query Likelihood*.

valores de correlación que se obtienen al aplicar este método de predicción para estimar la calidad del conjunto de documentos devueltos. Se aplican las funciones de ranking *BM25* y *Query Likelihood* sobre el conjunto de las tres colecciones de referencia en esta tesis, *TREC Vol. 4+5* (Tabla 3.5), *WT10g* (Tabla 3.6), y *GOV2* (Tabla 3.7).

De los resultados que se obtienen con esta primera aproximación se pueden extraer importantes conclusiones. Primero, es destacable que en general aparece un grado de correlación significativo, independientemente de la colección o función de ranking donde se evalúe el método de predicción. El grado de correlación aumenta de forma considerable, algo que es esperable, cuando se fija un número de documentos k adecuado. Por tanto, se comprueba el efecto negativo que produce la 'larga-cola' al estimar la calidad de las consultas vía desviación.

Si comparamos la calidad de las estimaciones por colección, destacan positivamente los resultados obtenidos con el conjunto de consultas de la tarea *Robust2004*. El grado de correlación para el conjunto de estas 249 consultas supera en el mejor de los casos $r = 0{,}6$, valor que es similar a aquellos que obtienen métodos de predicción bastante más complejos.

En la parte negativa se encuadran los resultados en la colección $WT10g$, donde en el peor de los casos aparecen valores de correlación no significativos,

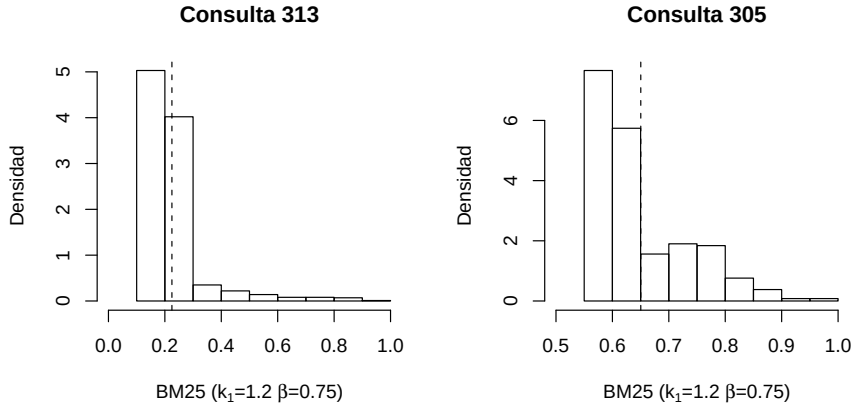

Figura 3.8: Histograma de los valores de relevancia que obtienen las consultas 313 y 305 (*Robust2004*). Los valores están normalizados por el máximo de cada consulta. El numero de documentos recuperados está fijado a 1000.

es decir no existe correlación. El valor de correlación más elevado sobre $WT10g$ apenas supera $r = 0{,}3$. Este comportamiento se analizará en la Sección 3.6.

Los valores de correlación encontrados en la colección $GOV2$ aparecen enmarcados entre los dos anteriores con valores de correlación r en el entorno de 0,45.

Los valores de correlación con *Kendall* son considerablemente menores a los obtenidos con *Pearson*, algo bastante común ya que en general *Kendall* produce valores de correlación comparativamente menores a los producidos al usar *Pearson*.

Finalmente, no aparece una tendencia clara respecto a si es más fácil predecir en base a los valores computados por *BM25* o *Query Likelihood*. Así, mientras *Query Likelihood* parece mostrarse menos sensible a cambios en el punto de corte k establecido, los valores de correlación r con *BM25*, al ser optimizado k, superan, en algunos casos, levemente a los obtenidos con *Query Likelihood*. Aparece también una muy leve tendencia, a tener estimaciones de mayor calidad por parte de *Query Likelihood*, si éstas son evaluadas con *Kendall*. Sin embargo, en base a los valores que se obtienen con *Pearson*, el comportamiento observado es el contrario.

En todos los resultados mostrados, aquellos valores de correlación que aparecen acompañados de un símbolo de asterisco, son estadísticamente significativos respecto a que el valor de correlación calculado no sea igual a cero con un $p - valor < 0{,}05$.

A través de este análisis se puede concluir que en general, los resultados que se obtienen con la simple utilización de un punto de corte común,

Tabla 3.5: Los valores de relevancia se normalizan en base al máximo valor de relevancia encontrado. σ_{total} corresponde a no establecer un punto de corte, por tanto se calcula la desviación para el total de la lista de documentos que en esta tarea correspondía a mil documentos. σ_{100} corresponde a establecer el punto de corte $k = 100$, donde se maximiza la correlación para ambos modelos de recuperación.

	BM25		**QL**	
	Pearson	**Kendall**	**Pearson**	**Kendall**
σ_{total}	0.2986^*	0.1905^*	0.4138^*	0.3303^*
σ_{100}	0.6058^*	0.3844^*	0.5429^*	0.3974^*

Tabla 3.6: Los valores de relevancia se normalizan en base al máximo valor de relevancia encontrado. σ_{total} corresponde a no establecer un punto de corte, por tanto se calcula la desviación para el total de la lista de documentos que en esta tarea correspondía a mil documentos. σ_{200} corresponde a establecer el punto de corte $k = 200$, donde se maximiza la correlación para ambos modelos de recuperación.

	BM25		**QL**	
	Pearson	**Kendall**	**Pearson**	**Kendall**
σ_{total}	0.2054^*	0.2137^*	0.1187	0.1772
σ_{200}	0.305^*	0.2006^*	0.2741^*	0.2163^*

que maximice la correlación, son bastante positivos. Sin embargo, esta aproximación puede ser problemática. Presenta, la dificultad de realizar una optimización de un conjunto de consultas para computar el valor k más adecuado globalmente para el conjunto de las consultas. Es decir, que se establezca un k común para todas las consultas, obviando de alguna manera las características específicas de cada consulta.

A priori, parece más adecuado que dicho punto de corte no se establezca para el conjunto de consultas, sino en base a cada una de las consultas. Es decir, para un conjunto de m consultas, se deben fijar m puntos de corte distintos k_i, de forma específica para cada consulta q_i.

En la siguiente sección, se estudia la utilidad y viabilidad de la estimación de un k_i dependiente para cada consulta. Para esta tarea, se plantea la aplicación de un método de optimización, a partir del cuál se pueda observar experimentalmente el teórico potencial incremento en términos de correlación de esta última aproximación. Además, se analizará si los distintos valores k_i de corte muestran cierta similitud o si por el contrario aparece un alto grado de variabilidad entre ellos.

Tabla 3.7: Los valores de relevancia se normalizan en base al máximo valor de relevancia encontrado. σ_{total} corresponde a no establecer un punto de corte, por tanto se calcula la desviación para el total de la lista de documentos que en esta tarea correspondía a diez mil documentos. σ_{2000} corresponde a establecer el punto de corte $k = 2000$, donde se maximiza la correlación para ambos modelos de recuperación.

	BM25		**QL**	
	Pearson	**Kendall**	**Pearson**	**Kendall**
σ_{total}	0.3425^{*}	0.2397^{*}	0.3867^{*}	0.3050^{*}
σ_{2000}	0.4593^{*}	0.2841^{*}	0.4471^{*}	0.3444^{*}

3.3.2. Optimización del punto de corte por consulta

Con el objetivo de analizar la potencial influencia, de fijar un punto de corte de forma específica para cada consulta, se realiza el siguiente estudio. El objetivo es fijar de forma óptima los distintos k_i, de forma que se maximice el grado de correlación entre la predicción obtenida y la precisión media de las consultas. Como resultado de la optimización se obtiene un umbral superior de correlación, que describe la utilidad de la estimación de distintos k_i. A su vez y en base a los valores k_i resultantes será posible evaluar la viabilidad de estimar dichos puntos de corte de forma automática.

La experimentación que se presenta a continuación se ha realizado sobre las colecciones *TREC Vol. 4+5* y *WT10g*, excluyendo en este caso a la colección *GOV2*. Ambas colecciones sirven al objetivo propuesto, ya que se han revelado como aquellas en las que mayor y menor nivel de correlación aparece respectivamente, según los resultados presentados en la Sección 3.3.1.

El proceso de optimización se basa en el uso de algoritmos genéticos utilizando como función de ajuste aquella que maximiza el valor de correlación.

Descripción del experimento

El algoritmo se diseña según las siguientes especificaciones. Cada elemento de la población o solución candidata es un vector de n posiciones, siendo n el número de consultas en la colección a optimizar, 249 para *TREC Vol. 4+5* y 97 para *WT10g*.

Cada posición del vector o gen representa el punto de corte para una consulta. El punto de corte de cada posición del vector se inicializa aleatoriamente con un valor entre 30 y 500.

Las mutaciones de carácter aleatorio desplazan el punto de corte un máximo de 300 posiciones. La dirección de dicho desplazamiento se elige también de forma aleatoria. Se añade la restricción de que el resultado del desplazamiento produzca un punto de corte válido, esto es, entre 10 y 1000.

Además, se aplica el operador de cruce monopunto, es decir dos elementos de la población generan un nuevo miembro de la población aportando cada uno la mitad de los cromosomas del nuevo elemento. En la Tabla 3.8 aparecen los parámetros específicos de dicho algoritmo.

Es importante destacar que se realiza un proceso de optimización por cada colección, función de ranking y coeficiente de correlación, lo que da como resultado un total de ocho procesos de optimización ($Col\ X\ Func\ X\ Corr$).

Tabla 3.8: Parámetros del algoritmo evolutivo de selección óptima de puntos de corte.

Parámetro	Valor
Longitud de cruce	1000
Población	100
Ratio de cruce	0.4
Ratio de mutación	0.1
Número máx. evaluaciones	10000

Los resultados de la experimentación realizada con las 249 consultas correspondientes a la tarea *Robust2004* aparecen en la Tabla 3.9. Se observa un importante incremento en el grado de correlación para cada uno de los casos, obteniéndose valores para el coeficiente *Pearson* en el entorno de 0,8 y para *Kendall* de 0,7. Como con la experimentación previa, se observa que la función de ranking *BM25* es más sensible al punto de corte, ya que el incremento en la correlación supera al que aparece con *Query Likelihood*. También se corrobora el hecho de que con el uso de *Kendall*, se obtienen valores de correlación levemente inferiores a los de *Pearson*.

Tabla 3.9: Los valores de relevancia son previamente normalizados utilizando el máximo encontrado por cada consulta.

TREC Vol. 4+5	**BM25**	**QL**
Pearson	0.9023	0.7979
Kendall	0.7951	0.6634

En el caso de la correlación obtenida con la colección *WT10g* (Tabla 3.10), se puede observar un empeoramiento general del grado de correlación respecto a *TREC Vol. 4+5*, algo que ya ocurría usando un punto de corte global para todas las consultas. De nuevo se corrobora lo observado en los resultados de *TREC Vol. 4+5*. Es decir, el rendimiento es menor según *Kendall* y se observa una capacidad de predicción menor usando *Query Likelihood* que *BM25*.

Aún así y en base al incremento en términos de correlación que aparece con ambas colecciones, se corrobora la posibilidad de mejorar el rendimiento

Tabla 3.10: Los valores de relevancia son previamente normalizados utilizando el máximo encontrado por cada consulta.

WT10g	**BM25**	**QL**
Pearson	0.7121	0.5785
Kendall	0.6914	0.5475

de las predicciones, al establecer un punto de corte específico por consulta.

Una vez observada la utilidad de fijar puntos de corte de manera específica, se debe analizar la viabilidad de estimar dicho punto de corte de manera automática. La dificultad de esta estimación puede evaluarse en base a el grado de variación encontrado entre los distintos k_i calculados con el método de optimización. Esta información aparece reflejada en la Tabla 3.11, para el caso de *TREC Vol. 4+5* y en la Tabla 3.12 para *WT10g*. En ambas tablas se muestra la media y la desviación de los distintos puntos de corte resultado de la optimización respecto a *Pearson* y *Kendall*.

Independientemente de la colección y del método de correlación a optimizar, se observa que para el caso de *BM25* el punto de corte se fija en el entorno de 350. Sin embargo, para *Query Likelihood* dicho punto de corte promedio se sitúa en valores más cercanos a 270. Aunque el valor de corte promedio que se obtiene permanece relativamente estable independientemente de la colección, el grado de desviación es muy elevado, superando ampliamente 300. Este hecho indica un muy elevado grado de variabilidad entre los puntos de corte óptimos, lo que incrementa la dificultad de establecer mecanismos automáticos para su cálculo.

Tabla 3.11: Media y desviación de los puntos de corte obtenidos para la colección *TREC Vol. 4+5*.

TREC Vol. 4+5	**Pearson**		**Kendall**	
	BM25	**QL**	**BM25**	**QL**
Media(k_i)	312.96	290.39	332.84	278.52
Desv(k_i)	339.72	366.23	342.85	329.21

Como consecuencia de este estudio experimental se proponen dos nuevos métodos para calcular el punto de corte específicamente por consulta. El objetivo principal de ambas aproximaciones es el de mejorar los resultados obtenidos con el uso de un punto de corte común para todas las consultas.

Tabla 3.12: Media y desviación de los puntos de corte obtenidos para la colección *WT10g*.

WT10g	**Pearson**		**Kendall**	
	BM25	**QL**	**BM25**	**QL**
Media(k_i)	339.45	234.02	373.75	249.50
Desv(k_i)	385.90	350.07	370.79	333.56

3.4. Estimación de un punto de corte por consulta

3.4.1. Desviación estándar máxima

El primer método propuesto para la estimación de los puntos de corte específicos para cada consulta, utiliza como heurística la propia evolución de la medida de la desviación estándar, a medida que se añaden nuevos documentos para su cálculo.

En la Figura 3.9 aparece la evolución de la desviación estándar para las consultas 313 ($AP \approx 0{,}7$) y 305 ($AP < 0{,}01$) de la tarea *Robust2004*, que ya fueron utilizadas de ejemplo en la Sección 3.3.

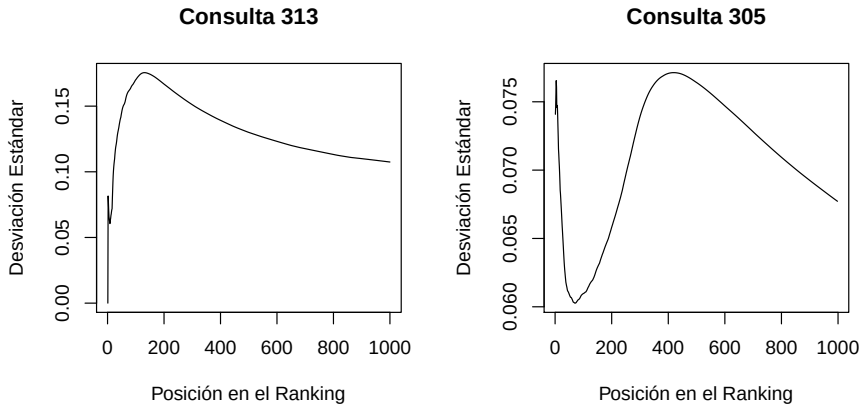

Figura 3.9: Los valores de relevancia se han normalizado en base al máximo valor de relevancia encontrado.

Como se observa en la figura, en general, la desviación estándar desciende de forma monótona una vez se ha alcanzado un máximo. Es decir, se alcanza un máximo que se va reduciendo de forma monótona según nuevos documentos, de supuestamente menor relevancia, van siendo recuperados.

La hipótesis de partida para la medida que se plantea a continuación es que estos documentos, que según la función de ranking son menos relevantes

para el usuario y disminuyen la desviación, son los que conforman de forma mayoritaria la denominada cola de la distribución exponencial.

Para reducir el efecto que introduce este conjunto de documentos, se plantea utilizar como punto de corte específico para cada consulta, aquel tamaño de lista donde la desviación se haga máxima. El cálculo de la desviación máxima se consigue midiendo la desviación estándar cada vez que se recupera un nuevo documento y seleccionando aquel punto donde ésta sea máxima. Con el objetivo de eliminar el efecto que se puede producir en el caso de que el primer documento del ranking obtenga un valor de relevancia muy elevado, la desviación estándar máxima se calcula para un tamaño de lista mínimo de cinco documentos.

La desviación máxima, σ_{max}, sobre un ranking de documentos, RL, que contiene N documentos d, se define de la siguiente forma:

$$\sigma_{max} = max[\sigma(RL_{[1,d]}) : d \in [5, N]]$$

A continuación se muestran los resultados que se obtienen a partir de esta aproximación, en comparación con la selección de un punto de corte común.

Aquellos valores de correlación que aparecen acompañados de un símbolo de asterisco, son estadísticamente significativos respecto a que el valor de correlación calculado no sea igual a cero con un $p - valor < 0{,}05$.

En el caso de la aplicación de la desviación máxima a las consultas de *Robust2004* (Tabla 3.13), se pueden observar grados de correlación inferiores para el caso *Pearson*. Esta perdida de rendimiento es de menor importancia al evaluar con *Kendall*. Este análisis es completamente extrapolable al caso de la colección *WT10g* (Tabla 3.14) y *GOV2* (Tabla 3.15). En estas dos últimas colecciones, aparecen casos en los que el coeficiente de correlación *Kendall* que se obtiene con el uso de la desviación máxima, supera ligeramente al que se obtenía con la selección de un punto de corte común.

Tabla 3.13: y σ_{100} para facilitar la comparación entre ambos métodos. Los valores de relevancia se han normalizado en base al máximo valor de relevancia encontrado.

	BM25		**QL**	
	Pearson	**Kendall**	**Pearson**	**Kendall**
σ_{100}	0.6058^{*}	0.3844^{*}	0.5429^{*}	0.3974^{*}
σ_{max}	0.5050^{*}	0.3784^{*}	0.4322^{*}	0.3632^{*}

Como resumen a la introducción de la medida de desviación máxima como método de predicción, se debe destacar que los resultados muestran un leve descenso en el rendimiento, algo que aparece de forma menos acusada al evaluar con *Kendall*. Sin embargo, el grado de correlación obtenido no

Tabla 3.14: y σ_{200} para facilitar la comparación entre ambos métodos. Los valores de relevancia se han normalizado en base al máximo valor de relevancia encontrado.

	BM25		**QL**	
	Pearson	**Kendall**	**Pearson**	**Kendall**
σ_{200}	0.305^{*}	0.2006^{*}	0.2741^{*}	0.2163^{*}
σ_{max}	0.2772^{*}	0.2270^{*}	0.2045^{*}	0.2210^{*}

Tabla 3.15: y σ_{2000} para facilitar la comparación entre ambos métodos. Los valores de relevancia se han normalizado en base al máximo valor de relevancia encontrado.

	BM25		**QL**	
	Pearson	**Kendall**	**Pearson**	**Kendall**
σ_{2000}	0.4593^{*}	0.2841^{*}	0.4471^{*}	0.3444^{*}
σ_{max}	0.4391^{*}	0.3400^{*}	0.3827^{*}	0.3111^{*}

se aleja en gran medida de lo mostrado con la selección de un punto de corte común. La principal ventaja de la aplicación de la desviación máxima es la selección automática de un punto de corte adecuado, que elimina la necesidad de optimizar k de forma global para el conjunto de consultas.

3.4.2. Aproximación basada en el operador *AND*

Otra aproximación que se propone en este trabajo, es la de aplicar el tamaño de la lista de documentos que se obtiene al utilizar el operador booleano *AND* en la consulta. El tamaño de la lista de documentos recuperada servirá como aproximación del tamaño ideal de la lista de resultados a la hora de aplicar la desviación sobre este conjunto de documentos. Al introducir el operador booleano *AND* en la consulta, solo aquellos documentos que contengan todos los términos de ésta serán recuperados. El operador *AND* suele producir el efecto de incrementar la precisión de la respuesta, con el coste de reducir en gran medida la cobertura de la consulta.

Esta propiedad se ajusta perfectamente al objetivo de eliminar de la forma más exacta posible, la denominada cola de la distribución exponencial. Aumentar la precisión supone reducir el número de documentos no relevantes recuperados, mientras que reducir la cobertura de una consulta supone dejar de recuperar documentos relevantes. Al reducir el número de documentos no relevantes recuperados, disminuye el tamaño de la cola de la distribución exponencial, ya que teóricamente ésta se compone en su mayoría de este tipo de documentos.

La aplicación del operador *AND* como estimador ha sido utilizado fre-

cuentemente con anterioridad. Por ejemplo, Hauff et al. (2008b), en su propuesta de mejora para *Clarity Score*, lo utiliza para aumentar la precisión del conjunto de documentos recuperados, que posteriormente utiliza para construir un modelo de lenguaje representativo de la consulta.

Otro indicio que apoya la utilidad del operador *AND* como estimador del punto de corte, parte de los resultados obtenidos en la Sección 3.3.2. Los puntos de corte por consulta calculados con el uso de algoritmos evolutivos en esta sección, muestran un leve grado de correlación, $r = 0{,}2$, con los puntos de corte resultantes de aplicar el operador *AND* directamente. Este hecho podría ser indicativo de la posible utilidad de la utilización del operador *AND*, como estimador del punto de corte ideal de forma específica para cada consulta.

El problema que aparece con la utilización de dicho operador es que es demasiado restrictivo para algunas consultas. Esto provoca que en algunos casos ningún documento de la colección sea recuperado, ya que ninguno contiene todos los términos de la consulta. En otros casos, para consultas con un único término o con términos muy frecuentes en la colección, el número de documentos recuperados es muy elevado. Un número tan elevado de documentos hace que aparezca de nuevo el problema de la cola de la distribución exponencial.

Con el objetivo de evitar ambas situaciones, se plantea aplicar una función de escalado de tal forma que dado un punto de corte k obtenido inicialmente con el operador *AND*, se calcula un nuevo punto de corte k'.

Una restricción importante de la función de escalado lineal que se aplica, es que la media de los puntos de corte resultantes de su aplicación debe ser igual a la media de los puntos de corte que aparecía inicialmente.

Por tanto aquellos valores k que se alejaban mucho de la media, es decir se situaban muy cercanos a cero o al máximo tamaño de la lista de resultados, se escalan en nuevos valores k' más cercanos a la media original $\bar{k}$.

Esta transformación requiere de un parámetro libre λ, que especifica cuál es el valor máximo de documentos k'_{max}, a partir de los cuales es posible medir la desviación. Es decir el tamaño máximo de la lista de resultados.

Formalmente, se aplican las siguientes restricciones a la función de escalado:

$$\bar{k'} = \bar{k} \text{ y } k'_{max} = \lambda \bar{k}$$

donde k'_{max} es el tamaño máximo esperado de documentos recuperados para cualquier consulta de las tratadas, fijando dicho valor al número de documentos máximo admitido, siendo este 1000 para *Robust2004* y *WT10g* y 10000 para *GOV2* que coincide en el tamaño de la lista de resultados específico para cada colección en sus respectivas tareas dentro de *TREC*. Así la transformación lineal, basándose en las condiciones previas, se calcula de la siguiente forma:

$$k' = ak + b$$

donde a y b se calculan como:

$$b = (1 - a)\tilde{k}$$

$$a = \frac{(\lambda - 1)\bar{k}}{k_{max} - \bar{k}}$$

donde $\lambda \in (1, \frac{k_{max}}{k}]$.

Los resultados que se obtienen con la aplicación descrita para el cálculo del punto de corte escalado, en base al número de documentos devueltos usando el operador *AND* aparece en la Tablas 3.16, 3.17 y 3.18.

Los resultados en la colección *TREC Vol. 4+5* se muestran en la Tabla 3.16. En este caso, se observan ligeros incrementos generales en el grado de correlación.

De nuevo la experimentación realizada en la colección *WT10g* (Tabla 3.17), obtiene los resultados de correlación más bajos. Las predicciones sobre la función de ranking *BM25* en *WT10g* muestran cierto grado de mejoría, mientras que para el caso de *Query Likelihood*, se observa un descenso en la calidad de las predicciones.

Finalmente para el caso *GOV2* que aparece en la Tabla 3.18, los resultados se ven incrementados en casi todos los casos, salvo al utilizar *BM25* y *Kendall* como función de evaluación, donde aparece una leve disminución del grado de correlación.

Aquellos valores de correlación que aparecen acompañados de un símbolo de asterisco, son estadísticamente significativos respecto a que el valor de correlación calculado no sea igual a cero con un $p - valor < 0{,}05$.

Tabla 3.16: con el parámetro $\lambda = 5$. Además se incluye el valor que se obtuvo con σ_{100} para facilitar la comparación entre ambos métodos. Los valores de relevancia se han normalizado en base al máximo valor de relevancia encontrado.

	BM25		**QL**	
	Pearson	**Kendall**	**Pearson**	**Kendall**
σ_{100}	0.6058^{*}	0.3844^{*}	0.5429^{*}	0.3974^{*}
$\sigma_{k,\lambda=5}$	0.6128^{*}	0.4074^{*}	0.5602^{*}	0.3846^{*}

Aunque se puede destacar que en términos generales la aplicación de *AND* como estimador incrementa el grado de correlación, este incremento no supone una mejora significativa en la calidad de las predicciones. La colección *WT10g* vuelve a destacar como la colección donde más difícil es realizar predicciones de calidad, mientras las consultas de *Robust2004*, por el contrario, obtienen niveles de correlación elevados.

Tabla 3.17: con el parámetro $\lambda = 2$. Además se incluye el valor que se obtuvo con σ_{200} para facilitar la comparación entre ambos métodos. Los valores de relevancia se han normalizado en base al máximo valor de relevancia encontrado.

	BM25		**QL**	
	Pearson	**Kendall**	**Pearson**	**Kendall**
σ_{200}	0.305^{*}	0.2006^{*}	0.2741^{*}	0.2163^{*}
$\sigma_{k,\lambda=2}$	0.3096^{*}	0.2253^{*}	0.2346^{*}	0.1793

Tabla 3.18: con el parámetro $\lambda = 5$. Además se incluye el valor que se obtuvo con σ_{2000} para facilitar la comparación entre ambos métodos. Los valores de relevancia se han normalizado en base al máximo valor de relevancia encontrado.

	BM25		**QL**	
	Pearson	**Kendall**	**Pearson**	**Kendall**
σ_{max}	0.4391^{*}	0.3400^{*}	0.3827^{*}	0.3111^{*}
$\sigma_{k,\lambda=5}$	0.4675^{*}	0.3291^{*}	0.4263^{*}	0.3414^{*}

Se repite, en cierta forma, la tendencia general observada a lo largo de la experimentación con el conjunto de medidas propuestas.

3.5. Análisis Comparativo de los Resultados

A modo de resumen se presentan el conjunto de resultados obtenidos con las distintas aproximaciones propuestas. El objetivo de mostrar el conjunto total de resultados tiene un doble objetivo: facilitar la observación de las diferencias que suponen las distintas aproximaciones propuestas y resaltar la capacidad de predicción de la medida de la desviación para las distintas colecciones.

3.5.1. Resultados para *TREC Vol. 4+5*

Las consultas de la tarea *Robust2004* (Tabla 3.19), resueltas sobre la colección *TREC Vol. 4+5*, se han revelado como aquellas en la que más fácil parece predecir la calidad de las consultas. Este hecho no aparece únicamente en la propuesta presentada en esta tesis, sino que la mayoría de los métodos de predicción propuestos en otros trabajos, muestran en esta colección su mayor rendimiento (He y Ounis (2006); Hauff et al. (2008b); Zhao et al. (2008); Hauff (2010)).

El hecho de que los resultados en esta colección superen a los del resto

de las colecciones utilizadas en esta tesis, no se puede considerar casual. Dentro de la tarea *Robust2004*, uno de los objetivos era ordenar las consultas suministradas por su dificultad, es decir se trataba de estimar la dificultad relativa de las 250 consultas suministradas originalmente. Dentro de este conjunto de consultas se introdujeron un conjunto de cincuenta denominadas 'Difíciles' con el objetivo de facilitar su detección (Voorhees (2004)).

En relación al rendimiento de las aproximaciones fruto de esta tesis, destacan la medida de la desviación usando un punto de corte común, y la que hace uso del operador *AND* como estimador. Con la aplicación de ambas aproximaciones se obtienen valores de correlación elevados en comparación con los valores que se suelen obtener con otros métodos de predicción.

Como ya se ha mencionado anteriormente, y se repite para el conjunto de colecciones, el grado de correlación *Kendall* suele mostrar coeficientes menores a los que aparecen al evaluar con *Pearson*. En relación a los modelos de recuperación se ha observado la misma tendencia en todas las colecciones. En general *BM25* obtiene mejores resultados de correlación, como consecuencia de que la aplicación de distintos métodos de selección del punto de corte se muestran más efectivos en esta función de ranking. *Query Likelihood* por su lado muestra un comportamiento más estable, independientemente del punto de corte establecido.

Tabla 3.19: Los valores de relevancia se han normalizado en base al máximo valor de relevancia encontrado.

	BM25		**QL**	
	Pearson	**Kendall**	**Pearson**	**Kendall**
σ_{total}	0.2986^{*}	0.1905^{*}	0.4138^{*}	0.3303^{*}
σ_{100}	0.6058^{*}	0.3844^{*}	0.5429^{*}	0.3974^{*}
σ_{max}	0.5050^{*}	0.3784^{*}	0.4322^{*}	0.3632^{*}
$\sigma_{k,\lambda=5}$	0.6128^{*}	0.4074^{*}	0.5602^{*}	0.3846^{*}

3.5.2. Resultados para *WT10g*

La calidad de las predicciones en la colección *WT10g* (Tabla 3.20), se reduce enormemente en comparación con las otras colecciones. Al contrario de lo que ocurre con *TREC Vol. 4+5*, en general los distintos métodos de predicción obtienen valores de correlación bajos al ser evaluados con *WT10g*. Posteriormente, en la Sección 3.6, se desarrolla un análisis que pretende dar explicación a este fenómeno.

En relación al comportamiento de la desviación como método de predicción en *WT10g*, de nuevo destacan ligeramente los métodos de punto de corte global y de uso del estimador *AND*. Se observan las mismas tendencias respecto a *Kendall* y *Query Likelihood* que aparecían en *TREC Vol.*

4+5. Sin embargo las diferencias que aparecen entre *Pearson* y *Kendall* o *BM25* y *Query Likelihood* son menores, simple consecuencia del hecho de la aparición de coeficientes de correlación muy bajos.

Tabla 3.20: Los valores de relevancia se han normalizado en base al máximo valor de relevancia encontrado.

	BM25		**QL**	
	Pearson	**Kendall**	**Pearson**	**Kendall**
σ_{total}	0.2054^{*}	0.2137^{*}	0.1187	0.1772
σ_{200}	0.305^{*}	0.2006^{*}	0.2741^{*}	0.2163^{*}
σ_{max}	0.2772^{*}	0.2270^{*}	0.2045^{*}	0.2210^{*}
$\sigma_{k,\lambda=2}$	0.3096^{*}	0.2253^{*}	0.2346^{*}	0.1793

3.5.3. Resultados para *GOV2*

Los resultados que se obtienen en *GOV2* (Tabla 3.21), se encuentran a medio camino de los que se mostraron en las dos colecciones anteriores.

Existen ciertas características propias de esta colección que pueden afectar al rendimiento mostrado por el método de predicción propuesto. Así, el gran tamaño de esta colección aumenta la probabilidad de aparición de documentos similares, lo que provoca a su vez, la recuperación de documentos con valores de relevancia muy parecidos. Este hecho reduce el grado de desviación estándar, lo que puede penalizar al método aquí propuesto.

Este efecto podría reducirse, al menos parcialmente, con el uso de funciones de ranking que utilizaran indicios de relevancia, no solo basados en estadísticos de los términos de la consulta. Por ejemplo, añadiendo información del grafo de enlaces entre los documentos como criterio de relevancia.

Como muestran los resultados se repite lo observado para el resto de colecciones.

Tabla 3.21: Los valores de relevancia se han normalizado en base al máximo valor de relevancia encontrado.

	BM25		**QL**	
	Pearson	**Kendall**	**Pearson**	**Kendall**
σ_{total}	0.3425^{*}	0.2397^{*}	0.3867^{*}	0.3050^{*}
σ_{2000}	0.4593^{*}	0.2841^{*}	0.4471^{*}	0.3444^{*}
σ_{max}	0.439^{*}	0.3400^{*}	0.3827^{*}	0.3111^{*}
$\sigma_{k,\lambda=5}$	0.4675^{*}	0.3291^{*}	0.4761^{*}	0.3443^{*}

Análisis comparativo

Finalmente y con el objetivo de comprobar el rendimiento de la desviación como método de predicción de forma comparativa, se muestran los resultados que se obtienen en relación a los métodos *Clarity Score* (CS) e *Improved Clarity Score* (ICS). Ambos métodos se describieron en profundidad en la Sección 2.4, y destacan por el rendimiento global que muestran en cuanto a la calidad de sus predicciones.

En el caso de las tablas que aparecen a continuación, las consultas se agrupan en grupos de cincuenta para facilitar la comparación con los resultados publicados previamente en Hauff (2010), pág. 83,84.

Los resultados que se han obtenido para la colección *TREC Vol. 4+5* aparecen en la Tabla 3.19, para *WT10g* en la Tabla 3.20 y los datos relativos a la colección *GOV2* se muestran en la Tabla 3.21.

En base a los resultados mostrados, se observa que la aproximación basada en el operador *AND*, obtiene resultados que se pueden considerar similares a los obtenidos por *CS* y su versión mejorada. Este hecho corrobora la idoneidad de la aproximación aquí presentada, ya que con la aplicación de técnicas simples sobre los valores de relevancia, se ha conseguido un rendimiento similar al de métodos que aplican técnicas más complejas.

Tabla 3.22: Los mejores resultados aparecen en negrita.

	Pearson			**Kendall**		
	301-350	**351-400**	**401-450**	**301-350**	**351-400**	**401-450**
$\sigma_{k,\lambda=5}(BM25)$	**0.8181***	0.5115*	0.6246*	0.5266*	0.3760*	0.3043*
$\sigma_{k,\lambda=5}(QL)$	0.5388*	**0.5333***	**0.7561***	0.3877*	0.3741*	0.4985*
CS	0.5390*	0.3095*	0.5727*	**0.6033***	0.4441*	0.4198*
ICS	0.6330*	0.5106*	0.7064*	0.5422*	**0.5498***	**0.4998***

Tabla 3.23: Los mejores resultados aparecen en negrita.

	Pearson		**Kendall**	
	451-500	**501-550**	**451-500**	**501-550**
$\sigma_{k,\lambda=2}(BM25)$	0.3100*	**0.3284***	0.2498*	0.1748
$\sigma_{k,\lambda=2}(QL)$	0.2372*	0.2786*	0.1567	**0.2913***
CS	0.2607*	0.2517*	0.1292	0.2233*
ICS	**0.5921***	0.2813*	**0.3748***	0.1843

3.6. Entropía entre los resultados de *WT10g*

En general los resultados de correlación obtenidos sobre el conjunto de colecciones que han formado parte del análisis aquí realizado, indican una correlación significativa salvo para el caso de la colección *WT10g*. Un aspecto interesante a analizar serían las razones que hacen que el rendimiento de los

Tabla 3.24: Los mejores resultados aparecen en negrita.

	Pearson			**Kendall**		
	701-750	**751-800**	**801-850**	**701-750**	**751-800**	**801-850**
$\sigma_{k,\lambda=5}(BM25)$	0.4175*	0.4522*	**0.5067***	0.2702*	0.3453*	**0.3571***
$\sigma_{k,\lambda=5}(QL)$	0.5265*	0.4342*	0.4896*	**0.4155***	0.3518*	0.2908*
CS	**0.6033***	0.4441*	0.3872*	0.4149*	0.3299*	0.2350*
ICS	0.5422*	**0.5498***	0.4289*	0.3723*	**0.4181***	0.2497*

distintos métodos de predicción disminuyan en gran medida su efectividad para esta colección.

Algunas de las razones que causan el bajo rendimiento observado han sido mencionadas previamente por Hauff et al. (2008b). Aquí, se destacan dos razones principales que dificultan la predicción en *WT10g*:

- Calidad de la colección. Una mirada detallada a los documentos de la colección *WT10g* permite observar que un número significativo de documentos contienen pocos o ningún término. Además con cierta frecuencia se encuentran documentos repetidos, junto con alguna sospecha acerca del contenido malicioso de alguno de los documentos en forma de aplicación de ciertas técnicas de *WebSpam*.

- Las características entrópicas de la colección, que suponen una clara desventaja para aquellos métodos de predicción basados en modelos de lenguaje, tales como *Clarity Score* o derivados.

  Si observamos las propiedades entrópicas de la colección *WT10g*, se revela un bajo valor de entropía en su vocabulario. Como consecuencia de este hecho, se apreciarán menores diferencias entre los distintos modelos de lenguaje que se estiman para realizar la predicción y por tanto un menor rendimiento de estos métodos.

  Esta similitud en el vocabulario de los documentos, puede ser causada por el uso de un estilo muy homogéneo a la hora de redactar contenidos en sitios web gubernamentales, que conforman la mayor fuente de documentos de la colección *WT10g*.

Parece claro que ambas razones pueden ser motivos de la perdida de rendimiento de la propuesta descrita en este trabajo. Sin embargo, una característica que no se ha analizado hasta ahora, es el efecto que producen en el rendimiento de los métodos de predicción, las diferencias en términos de precisión media que muestran las consultas.

A priori, se puede suponer que en aquellas colecciones en las que aparezcan diferencias más notables entre las consultas, en términos de precisión media, la predicción pueda ser una tarea más sencilla.

Por tanto, se parte de la hipótesis de que es más sencillo realizar predicciones adecuadas cuanto mayor sea la divergencia que muestren las consul-

tas, en relación a la función objetivo de calidad con la que se evalúan.

Para medir las diferencias que muestran las consultas en términos de precisión media, proponemos utilizar el grado de entropía que obtienen el total de consultas, para el conjunto de las colecciones utilizadas en este trabajo.

El cálculo de la entropía en una distribución de probabilidad discreta, es algo inmediato pero en este caso la distribución de probabilidad de los valores de precisión media es continua, ya que se encuentra definida entre 0 y 1. Esto es, el valor de $AP \in [0,1] \in \Re$.

La entropía de una distribución de probabilidad continúa se calcula en base a su función de densidad. Así, dada una variable aleatoria X con una función de densidad $f(x)$, que toma valores $\in \Re$, se define la entropía diferencial como:

$$H(f) = -\int f(x) ln f(x) dx \tag{3.1}$$

La ecuación 3.1 ya aparecía en la propuesta original desarrollada por Shannon (1948). El principal problema que surge a la hora de calcular la entropía de los valores de AP sobre las distintas colecciones, es que es necesario conocer la distribución de probabilidad que describe los valores obtenidos.

Para el resultado obtenido por un sistema específico, con un número n de consultas, la distribución de probabilidad vendrá dada por una distribución multivariable. Así, cada consulta será definida por una distribución de probabilidad beta con parámetros α y β desconocidos, siendo la media de la distribución $\frac{\alpha}{\alpha+\beta}$ y la varianza $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Es decir, se tiene una distribución *Dirichlet* compuesta por n betas.

El ajuste de una distribución *Dirichlet* compuesta por n betas, es muy costoso y probablemente implicaría asumir un grado de error importante. Esta complejidad crece de manera exponencial, al tener en cuenta que el análisis aquí propuesto no se realiza sobre un único sistema. Con el objetivo de dar la máxima validez a los valores de entropía calculados, el estudio se realiza sobre un conjunto de sistemas que han participado a lo largo de los años en tareas que hacen uso de las consultas analizadas.

Por tanto, se opta por realizar una aproximación al valor de la entropía, en base a un muestreo calculado sobre los valores que obtienen los distintos sistemas que forman parte de este estudio. La aproximación basada en muestreo se basa en el trabajo de Beirlant et al. (1997), donde el objetivo es aproximar la función de densidad de la distribución de la probabilidad objetivo, en base al análisis de las muestras disponibles por intervalos.

Así, la aproximación basada en muestras se define por intervalos según la siguiente ecuación:

$$H_{m,n} = \frac{1}{n}\sum_{i=1}^{n-m} ln\left(\frac{n}{m}(X_{n,i+m} - X_{n,i})\right) - \psi(m) + ln\ m \tag{3.2}$$

donde $m = 1$ e indica el tamaño de los intervalos, $\psi(1) = -0{,}5772157$ que es el valor de la función de probabilidad digamma para $m = 1$, suponiendo la existencia de n variables aleatorias y una muestra de estas $X_1, X_2, ..., X_n$.

Debe tenerse en cuenta que el cálculo de la entropía diferencial para funciones de densidad f definidas en el espacio $[0, 1]$ tiene su máximo en cero. Es decir, para el caso de una distribución uniforme continua definida en $[0, 1]$, $H(f) = 0$. Por tanto las estimaciones de entropía que se obtengan sobre los distintos valores de *MAP* serán negativas, y tendrán un mayor valor entrópico cuanto más se acerquen a cero.

Los valores de entropía diferencial y el número de sistemas participantes en el estudio aparecen en la Tabla 3.25. La entropía de cada sistema se calcula de forma separada, calculándose finalmente la entropía promedio entre todos los sistemas.

Los resultados obtenidos concuerdan con lo esperado y con el rendimiento mostrado por los métodos de predicción propuestos en esta tesis. Así, de forma general aparece un significativo valor inferior de entropía para la colección *WT10g*, siendo la entropía en *Robust2004* claramente la mayor, lo que conlleva el menor rendimiento mostrado para *WT10g*.

Tabla 3.25: Valores de entropía entre los valores de precisión media para los sistemas que participaron en *TREC Vol. 4+5*, *WT10g* y *GOV2*.

	TREC Vol. 4+5	**WT10g**	**GOV2**
Entropía	-0.29	-0.63	-0.39
Sistemas	110	97	61

3.7. Conclusiones

En este capítulo se ha mostrado la utilidad de la aplicación de la desviación estándar sobre los valores de relevancia que computa una función de ranking, para estimar la calidad de respuesta que obtendrá una consulta. Las principales aportaciones se pueden resumir en los siguientes puntos:

- Desarrollo de una nueva propuesta de método de predicción, basado en el cálculo de estadísticos simples que pueden ser calculados sin incurrir en un elevado coste computacional.
- La calidad de las estimaciones del método propuesto son similares a las obtenidas con métodos más complejos.
- Se ha analizado el efecto que tiene el proceso de normalización de los valores de relevancia en relación a la predicción.

- Se ha comprobado el efecto que produce la cola de la distribución exponencial, que corresponde de forma teórica con los documentos no relevantes, a la hora de calcular las predicciones.
- Se han realizado distintas propuestas con el objetivo de reducir, el efecto negativo provocado por dicha cola.
- Se ha establecido un umbral máximo en la calidad de las estimaciones, que de forma teórica podrían ser alcanzadas por el método propuesto.
- Finalmente debido a los pobres resultados obtenidos en la colección *WT10g*, se ha analizado la relación entre la calidad de las predicciones y la variabilidad que presentan las consultas en términos de precisión media. Como consecuencia, se ha comprobado experimentalmente que la existencia de un mayor grado de variación en los resultados que obtienen las consultas, favorece la posibilidad de realizar predicciones de calidad.

I believe there is a direct correlation between love and laughter.

Yakov Smirnoff.

Capítulo 4

Evaluación de los Métodos de Predicción

4.1. Introducción

El gran interés surgido en el ámbito de la predicción de la calidad de consultas no ha venido acompañado de un estudio completo y profundo del marco de evaluación que se aplica en este campo. Algunos autores han destacado previamente, algunas de las limitaciones del enfoque actual de evaluación basado en distintos coeficientes de correlación. Estas limitaciones, junto a la no existencia de alternativas motivan claramente la necesidad del desarrollo de un nuevo enfoque de evaluación en el campo de la predicción.

En este capítulo se presenta un análisis crítico a la metodología de evaluación de los métodos de predicción, que reflejará las principales causas por las cuales se considera necesario un nuevo marco de evaluación. En base a este análisis se introduce una nueva propuesta de evaluación, cuyo principal objetivo es subrayar las diferencias de rendimiento que muestran distintos métodos de predicción. Esta propuesta de marco de evaluación es una de las principales aportaciones de esta tesis.

El contenido de este capítulo se estructura de la siguiente forma: en la primera sección se realizará una introducción a la metodología de evaluación que se aplica en la actualidad, con un especial énfasis en destacar algunas de las principales carencias del enfoque basado en coeficientes de correlación. A continuación, se describe el nuevo marco de evaluación propuesto en esta tesis con el objetivo de superar algunas de las limitaciones descritas con anterioridad. Finalmente, se presentan y comparan los resultados obtenidos utilizando ambos marcos de evaluación en una colección estándar de *TREC*. El análisis de los resultados que se obtienen, permite comprobar que con el uso del nuevo marco de evaluación, se proporciona mayor grado de detalle en relación al rendimiento de los métodos de predicción.

4.1.1. Coeficientes de correlación

Como se ha mencionado previamente en esta tesis, el rendimiento de los métodos de predicción se evalúa en términos de la correlación que muestran las predicciones, respecto a una medida de calidad de la respuesta. Generalmente esta medida se corresponde con la precisión media, que es considerada como estándar dentro del paradigma *Cranfield*. De esta forma se considera que un método de predicción es más preciso cuanto mayor correlación muestre con la precisión media.

El grado de correlación entre dos variables aleatorias mide la relación entre ambas variables. Dicha relación se cuantifica con un valor entre $[-1, 1]$, donde 1 implica una correlación perfecta directa, -1 significa que existe una relación perfecta pero de carácter inverso y 0 implica que no se ha observado ningún tipo de relación entre ambas variables.

Por tanto, para valores de correlación cercanos a cero, se puede decir que no se observa dependencia entre ambas variables aunque este hecho no se puede interpretar como que ambas variables sean completamente independientes entre sí.

De la misma manera, la observación de un alto grado de correlación negativo o positivo entre dos variables, sugiere pero no asegura, la existencia de una posible dependencia entre ambas.

Una de las críticas clásicas a la evaluación de métodos de predicción basada en correlación es consecuencia de que este tipo de evaluación en muchas ocasiones produce resultados muy similares para distintos métodos. Esto que dificulta en gran medida la interpretación de los resultados, ocultando las características específicas de cada uno de los métodos de predicción evaluados (Hauff (2010), pág. 148,149).

Otra de las carencias que presenta el marco de evaluación actual es su falta de orientación a aplicaciones específicas. Un marco de evaluación debería ayudarnos a decidir si el método de predicción evaluado, es adecuado para un contexto específico. El tipo de evaluación que se aplica en la actualidad proporciona un resultado muy general en relación al grado de acierto del método de predicción, que no permite particularizarse para grupos específicos de consultas. Por ejemplo, un marco de evaluación enfocado a la aplicación de métodos de predicción en un entorno concreto, debería tener la capacidad de responder cuestiones como: *¿Es el método de predicción A capaz de superar al método B en relación a la detección de consultas consideradas difíciles?*

Para poder dar respuesta a la pregunta anterior, es necesario desarrollar un nuevo marco de evaluación capaz de hacer explícitas las diferencias en rendimiento, cuando los métodos de predicción se aplican a consultas de distintos niveles de dificultad.

Este objetivo puede ser alcanzado asumiendo una clasificación discreta de las consultas, es decir, asumiendo que cada una de las consultas pertenece

a un tipo único basado en la calidad de la respuesta que obtenga dicha consulta. Por ejemplo, se puede considerar como consultas fáciles a aquellas que obtienen valores elevados de precisión media, frente a las difíciles con valores muy bajos.

Partiendo de esta hipótesis de trabajo es posible evitar algunas de las limitaciones que aparecen al aplicar la evaluación basada en coeficientes de correlación. La aproximación propuesta, permitirá evaluar los métodos de predicción con dos criterios: medir el rendimiento globalmente, de forma similar a como es realizado por los coeficientes de correlación, y en base al rendimiento que muestren frente a tipos específicos de consultas.

4.2. Marco actual de evaluación

La evaluación de los distintos métodos de predicción del rendimiento de consultas se realiza a través del uso de coeficientes de correlación, siendo *Pearson* (r) y *Kendall* (τ) los dos métodos más utilizados habitualmente [1].

Pearson es un método paramétrico que supone la existencia de una relación lineal entre las dos series de datos de las cuales se desea conocer su grado de correlación, indicando la dirección y grado de dicha correlación. Por otro lado *Kendall* calcula el grado de correlación entre dos series de datos en base a los intercambios entre elementos, que serían necesarios para que ambas series estuvieran ordenadas de forma equivalente. *Kendall* es un método no paramétrico y por tanto no realiza ninguna suposición sobre los datos. En general se considera a *Kendall* un método menos informativo aunque más robusto que *Pearson*.

Simplificando se podría decir que, mientras *Pearson* es un método enfocado a detectar si entre ambas series de datos existe algún tipo de relación lineal, *Kendall* mide la similitud entre el orden que muestran ambas series. Es decir, *Kendall*, en el contexto de la predicción del rendimiento de consultas, mide la similitud entre el ranking de según el valor de precisión media y de predicción.

Como consecuencia de las diferentes características que analizan ambos métodos de correlación, los resultados que se obtienen con ambos coeficientes no pueden ser comparados de forma directa. Por ejemplo en algunos casos los resultados pueden llevar a conclusiones contradictorias respecto a qué método de predicción es más acertado, ya que un método que muestre una muy buena correlación en términos de *Pearson*, no tiene porque mostrar el mismo grado de correlación en relación a *Kendall*.

[1] En algunos trabajos previos se ha utilizado el coeficiente de correlación *Spearman* (ρ) aunque de forma menos habitual.

4.2.1. Pearson

Las principales carencias que muestra la evaluación de métodos de predicción de rendimiento de consultas basada en *Pearson* han sido ampliamente tratadas con anterioridad en la literatura relacionada (Hauff et al. (2009)) o con el clásico cuarteto de *Anscombe* descrito en la Sección 2.4.3. Así, el coeficiente de correlación *Pearson* puede estimar de manera incorrecta la dependencia o relación entre dos series de datos, cuando esta relación se sustenta en una función no lineal. También es conocido el importante efecto que tiene sobre la medida final la presencia de datos atípicos, es decir datos muy alejados de la media muestral, por lo que se recomienda su exclusión del estudio antes de medir la correlación.

4.2.2. Kendall

Debido a algunas de las limitaciones que presenta *Pearson* es habitual que los resultados de la evaluación de un método de predicción se muestren en términos del coeficiente de correlación *Kendall*, asumiendo el inconveniente de usar una medida menos informativa.

Aunque se considera a la medida *Kendall* como una aproximación más adecuada para evaluar el rendimiento de distintos métodos de predicción, el uso de esta medida continúa ocultando algunos detalles del rendimiento de estos métodos.

Con el uso de *Kendall* se obtiene un único valor que representa el rendimiento para todo el conjunto de consultas, pero lo cierto es que en muchos casos el interés puede estar más centrado en un subconjunto específico de éstas. Por ejemplo, en aquellas consultas que compartan un nivel específico de calidad. En el computo del coeficiente τ cada uno de los elementos que pertenecen a las series de datos a evaluar, poseen una importancia equivalente independientemente de sus características específicas y por tanto se evalúa el orden relativo estricto de todo el conjunto de elementos.

En muchos casos obtener un conjunto de predicciones que permitan ordenar las consultas de forma equivalente a como se haría en base a la medida de calidad objetivo, no es estrictamente necesario para aplicar un método de predicción en un contexto específico.

Así, puede ocurrir que simplemente exista interés en observar el rendimiento en la predicción de aquellas consultas que superan cierto umbral de precisión media. Por tanto, no sería necesario evaluar el orden relativo para todo el conjunto de consultas.

La predicción de la calidad de la respuesta para consultas que obtienen un valor alto de precisión media es de vital importancia en áreas como la de la expansión de consultas, ya que nos puede ayudar a detectar en qué caso se hace recomendable llevar a cabo la expansión, como ha sido analizado por He et. al (He y Ounis (2009)).

El siguiente ejemplo es un caso típico donde se observa que la evaluación basada en el orden estricto de cada uno de los elementos ignora algunas propiedades interesantes del ranking, que resulta de aplicar un método de predicción.

Sean las siguientes series de datos $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ e $Y = \{\mathbf{4}, 2, 3, \mathbf{1}, 5, \mathbf{10}, 7, 8, 9, \mathbf{6}\}$, donde X representa las consultas ordenadas según su valor de precisión media, e Y representa el orden de las consultas según un método de predicción cualquiera.

De esta forma, según el valor real de calidad obtenido por las consultas, aquellas que podemos considerar 'fáciles' serían de la una a la cinco, mientras que las más difíciles serían de la seis a la diez y exactamente en este orden.

Puesto que los elementos 1,4,6 y 10 de la serie Y no están situados según el orden relativo de acuerdo a la serie de datos X, el valor de correlación *Kendall* que se obtiene corresponde a $\tau = 0{,}46$.

Este valor sugiere la existencia de cierta relación entre ambas series de datos. Sin embargo, si estuviéramos interesados en evaluar el orden existente por el tipo de consultas ("fáciles" o "difíciles") concluiríamos que en ambas series de datos, los elementos correspondientes a ambos tipos de consultas han sido clasificados de la misma forma.

Esto es, las consultas 'fáciles' y las 'difíciles' se encuentran agrupadas de igual forma en ambas series, ya que los elementos que pertenecen al grupo de las 'fáciles' se encuentran situadas en las primeras cinco posiciones, mientras que aquellas consultas que hemos considerado 'difíciles' ocupan las cinco últimas posiciones del ranking.

Imaginemos ahora el siguiente orden de consultas que se ha obtenido con otro método de predicción distinto $Y_2 = \{1, 2, \mathbf{6}, \mathbf{10}, \mathbf{3}, \mathbf{4}, \mathbf{5}, 7, 8, 9\}$.

Este nuevo método de predicción obtendría un valor de $\tau = 0{,}6$ bastante superior al anterior, por lo que deberíamos concluir que este segundo método de predicción es más apropiado que el primero.

Sin embargo mirando con detalle el nuevo orden establecido, podemos observar que el nuevo método de predicción considera como consultas 'fáciles' a la seis y a la diez y como 'difíciles' a la cuatro y a la cinco. Los errores que aparecen con este último método de predicción pueden perjudicar en mayor medida la calidad de nuestro sistema que aquellos que aparecen con el primer método de predicción. Por ejemplo, nuestro sistema de búsqueda al considerar que la consulta diez tiene una calidad alta, podría decidir suministrar al usuario un conjunto de resultados de muy baja calidad.

Una primera aproximación para intentar que el coeficiente de correlación *Kendall* detecte estas situaciones consiste en realizar mediciones parciales por cada grupo de elementos a considerar.

Esta primera aproximación aunque conceptualmente puede parecer adecuada, produce resultados indeseados ya que no tiene en cuenta el total de elementos que componen las series de datos de estudio, ignorando por tanto su orden relativo. Por ejemplo, si se considera la siguiente serie de

datos $Z = \{6, 7, 8, 9, 10, 1, 2, 3, 4, 5\}$, donde las 5 primeras posiciones corresponderían a las consultas consideradas 'fáciles' y el resto a las 'difíciles', según un nuevo método de predicción. El valor τ que se obtiene entre X y Z, usando ambos grupos por separado, sería en ambos casos igual a 1. Por el contrario, al observar ambas series de datos podemos observar que la predicción Z muestra un rendimiento muy pobre puesto que ha predicho aquellas consultas 'fáciles' como 'difíciles' y viceversa, esto es, el peor resultado posible.

Kendall con pesos

Debido a las limitaciones que muestra *Kendall* en relación a la evaluación de la predicción de consultas, algunos autores han planteado alternativas que se basan en asignar un grado de importancia a cada uno de los elementos sobre los que se mide la correlación τ. Así, algunos elementos contribuyen en mayor o menor medida a la medida final que se obtiene.

Este tipo de alternativas se suele denominar *Weighted Kendall* o *Kendall con Pesos* y se suele usar para medir la similitud entre rankings de distintos motores de búsqueda. Así, se considera que aquellos elementos que aparecen al principio de la lista de resultados, y por tanto tienen una mayor probabilidad de ser relevantes para el usuario, deben contribuir más a la medida final de correlación τ (Melucci (2009); Yilmaz et al. (2008)).

Aunque este tipo de evaluación puede ser más adecuada en el contexto de la predicción de la calidad de consultas, presenta dos claras desventajas.

En primer lugar se debe asignar un valor de contribución a cada uno de los elementos o consultas, lo cuál no es inmediato y tendrá un efecto muy importante en el valor final que obtengamos. Por otro lado, este tipo de aproximaciones no son capaces de suministrar resultados parciales según un tipo de consultas específico. Para realizar evaluaciones parciales es necesario evaluar todo el conjunto de predicciones modificando el valor de contribución de cada consulta y así enfocar la evaluación a cada tipo distinto de consulta.

4.3. Evaluación utilizando rangos de dificultad

Como se observó en la sección anterior, en la actualidad evaluamos los métodos de predicción en base a dos alternativas:

- *Pearson*: Capacidad de predecir el valor exacto de precisión media, o una función lineal de ésta, de un conjunto de consultas.

- *Kendall*: Capacidad para predecir el orden exacto, obviando el valor exacto de calidad de las consultas, si éstas fueran ordenadas según la calidad de su respuesta.

Aunque el objetivo de un método ideal de predicción sería estimar el valor exacto de calidad de cada una de las consultas, o en su defecto el orden exacto de éstas, ambos escenarios están bastante alejados del rendimiento actual que muestran las distintas técnicas de predicción. Sin embargo, se les aplica un un estándar de evaluación tan exigente que hace que ciertos detalles relativos al rendimiento de los métodos de predicción sean ignorados. Existen diversos ámbitos de aplicación de las predicciones, en los que no es estrictamente necesario ordenar perfectamente el conjunto de consultas, sino hacer explícitas las diferencias principales que se observan en dicho conjunto.

Por ejemplo, al estudiar el rendimiento de dos métodos de predicción como *Clarity Score* o el introducido en el Capítulo 3 de esta tesis basado en la desviación (*ScoreDesv*), se puede observar que en general su rendimiento aumenta a la hora de realizar predicciones sobre un conjunto de consultas entre las que existen diferencias de calidad importantes.

Para destacar este hecho se plantea dividir el conjunto de consultas de la tarea *Robust2004* en dos grupos distintos. Un primer grupo denominado 'Fáciles-Difíciles' de 100 consultas, formado por la unión de las 50 consultas que obtienen un mayor valor de precisión media, junto con las 50 consultas que obtienen un valor menor. El otro subconjunto denominado 'Promedio' corresponde a las 149 consultas restantes. Por tanto el grupo de consultas 'Fáciles-Difíciles' muestra mayores diferencias en lo que a precisión media se refiere en relación al subconjunto 'Promedio'.

A continuación se mide la capacidad de predicción de los métodos propuestos en estos dos conjuntos y en el total de consultas como suele ser habitual. Los resultados obtenidos aparecen en la Tabla 4.1. A partir de estos resultados es fácil concluir que, como era de esperar, los métodos de predicción muestran un mayor rendimiento cuanto más distintos son los elementos a predecir. Es decir, la predicción es de mayor calidad en el grupo de consultas 'Fáciles-Difíciles', disminuyendo el rendimiento de la predicción significativamente en el grupo denominado 'Promedio'. El rendimiento para el total de consultas se sitúa en medio del que obtienen los dos grupos anteriores.

Tabla 4.1: El conjunto total de consultas corresponde a las 249 consultas de la tarea *Robust2004*. La función de ranking aplicada es *Query Likelihood*. Los métodos de predicción corresponden a la versión clásica de *Clarity Score* y *ScoreDesv* utilizando 100 como punto de corte global.

	Clarity Score	ScoreDesv(100)
Fáciles-Difíciles	0.4768	0.5419
Promedio	0.3367	0.2456
Total	0.4184	0.3974

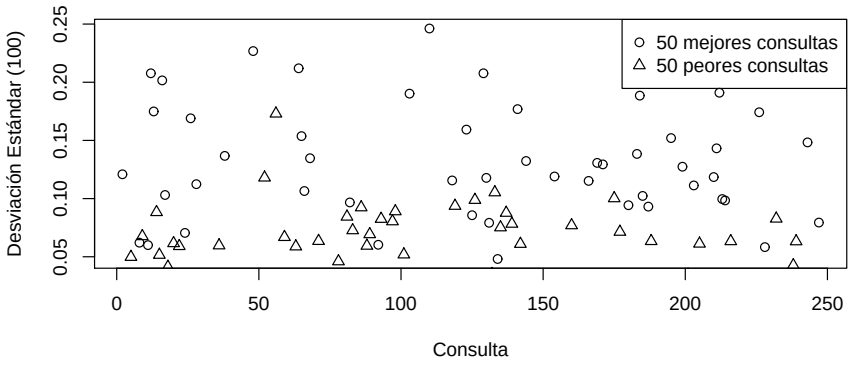

Figura 4.1: La función de ranking aplicada corresponde con *Query Likelihood*. El punto corte k para el método de predicción se estableció en 100.

La principal razón que explica estos resultados es que en general es más fácil predecir la existencia de grandes diferencias entre consultas. De la misma forma, es más fácil diferenciar entre el color que presentan un conjunto de bolas negras y blancas, que distinguir un conjunto de estas con distintos grados de grises.

De forma visual se puede observar esta misma situación en la Figura 4.1, donde aplicando como método de predicción la desviación estándar con $k = 100$, aparece una clara tendencia que divide el conjunto de consultas según el valor del método de predicción. Así, las mejores consultas, representadas por un círculo, se sitúan en la parte superior del gráfico, mientras las consultas con menor valor de precisión media, representadas con un triángulo, aparecen en la parte baja del gráfico.

Por tanto para tratar de evitar o reducir al máximo algunas de las limitaciones descritas, se plantea un nuevo marco de evaluación con dos objetivos principales:

- Evaluar no solo en base a un valor absoluto que describa el rendimiento global de un método de predicción, que en muchos casos dificulta su interpretación, sino también de manera específica para el marco de aplicación para el que se destina la predicción.

- Hacer explícito de manera cuantitativa el rendimiento para distintos grupos de consultas.

Un aspecto importante que debe ser precisado es qué se considera como una consulta 'fácil' o 'difícil'. Por convención estimamos que las consultas

fáciles son aquellas que obtienen un valor alto en términos de precisión media, mientras que las difíciles corresponderán a las consultas que muestren un bajo rendimiento en relación a la precisión media.

Para el desarrollo del nuevo marco de evaluación, debemos ser capaces de dividir el conjunto de m consultas en n bloques de interés, siendo $m \geq n$, donde cada consulta se asigna de forma única a una de las particiones en base al valor de precisión media que haya obtenido[2]. Así, se define el grado de calidad de cada una de las particiones como el valor promedio de los elementos que componen dicha partición.

Una vez ordenadas las consultas según su precisión media, las consultas de mayor calidad serán asignadas a la primera partición, para de forma equivalente definir las siguientes *n-ésimas* particiones. Este proceso finalizará cuando cada una de las consultas haya sido asignada a una de las particiones de manera única.

Este mismo proceso debe ser repetido pero en este caso utilizando el valor obtenido por el método de predicción aplicado, obteniendo así un número equivalente de particiones según ambos criterios.

Como resultado del doble proceso de partición del conjunto de consultas, a cada una de éstas se le asigna dos etiquetas, una es la partición a la que pertenezca en base a su calidad real y otra es la partición según el valor de predicción. De esta forma, si una consulta pertenece a dos particiones equivalentes, se considerará como un acierto en la predicción. Es inmediato reconocer que a través del doble proceso de partición, el problema de evaluación de métodos de predicción ha sido transformado en la evaluación de un problema de clasificación multiclase con n clases.

Grupos de consultas

Un problema interesante que surge con este nuevo modelo de evaluación es como agrupar las consultas en base a su rendimiento.

Un requisito previo que se debe exigir al método de agrupamiento, es que debe basarse principalmente en la distribución de datos generada por los valores de precisión media obtenidos por cada consulta, más allá de características específicas de cada consulta que puedan sugerir la dificultad de ésta, tales como la ambigüedad, la longitud o su semántica.

El problema de utilizar características específicas de una consulta, es que lo que para un sistema de recuperación puede ser una consulta de gran dificultad, quizá para otro pueda ser mucho más fácil debido a diversos factores, como podría ser la especialización de este último sistema en dar respuesta a consultas de la temática concreta que representa la consulta. Es más, con la simple observación del rendimiento de las consultas, se puede concluir que

[2] Aunque en general siempre se menciona como medida de calidad la precisión media, al igual que en la evaluación basada en correlación, es posible utilizar cualquier medida de calidad que se considere objetivo como por ejemplo *P@10*.

en algunos casos pudiera carecer de sentido la aplicación de un método de predicción. Esta importante conclusión se puede obtener mediante la simple observación de la distribución de datos generada por los valores de la medida de calidad en uso.

Por ejemplo, se puede imaginar un sistema en el cuál todas las consultas fueran respondidas de forma que satisfagan completamente al usuario, (precisión media muy elevada). En este caso extremo, el conjunto de consultas debería concurrir en una única partición ya que la calidad de las respuestas es muy similar. Algo equivalente ocurriría en el caso contrario, en el que el sistema de recuperación proporcionara respuestas muy pobres en todos los casos. En ambos escenarios la aplicación de un método de predicción carecería de sentido. Por tanto, la utilidad de un método de predicción cobra realmente sentido cuando las respuestas son heterogéneas en términos de calidad. Esta heterogeneidad debe ser explicitada por el método de agrupamiento, en base a las diferencias de rendimiento que aparezcan entre las consultas.

Por lo tanto, se considera adecuado que el método de agrupamiento sea completamente dirigido por los valores de la medida de calidad que caracteriza a cada una de las consultas.

Ninguna de los dos hipotéticos escenarios descritos previamente reflejan el rendimiento real de los motores de búsqueda que podemos encontrar en la actualidad. Es más, se puede comprobar que dentro del contexto de *TREC*, típicamente los valores de precisión media obtenidos para aquellas consultas de una tarea específica, suelen tener una distribución de datos que recuerda en gran medida a una distribución de probabilidad exponencial. Esta forma exponencial se va corrigiendo, tendiendo a algo más similar a una distribución normal a medida que la calidad general de la respuesta para todas las consultas aumenta. Este hecho se puede observar en la Figura 4.2. En esta figura aparecen los histogramas, junto con su función de densidad, de los valores de precisión media obtenidos en las tareas *Robust2004*, *WT10g* y *GOV2*. Las columnas representan el peor sistema, sistema promedio y el sistema ganador de la tarea, respectivamente. Así, en la primera fila aparece el peor, el promedio y el mejor sistema que participó en *Robust2004*. Como se puede comprobar de forma visual, la forma que dibujan la distribución de datos de los sistemas peores y promedios se asemeja a una distribución exponencial. En cambio aquellas distribuciones de datos obtenidas a partir de los sistemas ganadores, que aparecen en la tercera columna, tienden a parecerse a una distribución normal. La forma normal aparece especialmente con los resultados obtenidos por el ganador de la tarea *GOV2*, en la fila tres, columna tres, que coincide con aquel que obtiene un valor promedio de precisión media mayor.

Una vez destacada la representatividad que proporciona evaluar la calidad de cada consulta solo en base a la medida de calidad establecida, se podría argumentar que dicha representatividad sería específica solo del sis-

tema donde se obtuvo dicho rendimiento. Para dotar de mayor robustez al agrupamiento de consultas en base a su calidad, se plantean dos escenarios de evaluación diferentes, uno orientado a un sistema único y otro con características más generales.

- Sistema único: Este escenario simula la evaluación de los métodos de predicción, para un sistema único específico. Por tanto, el concepto de consulta 'fácil' o 'difícil' depende solo de la calidad que la consulta muestre en dicho sistema. Dicho escenario se correspondería con aquel en el que se desea comprobar la utilidad de incluir un método de predicción en un sistema de recuperación específico.

- Sistema genérico: Este escenario utiliza el valor promedio de calidad para cada consulta a partir de un conjunto de sistemas distintos. Así, al utilizar un conjunto de sistemas lo más heterogéneo posible, el concepto de consulta 'fácil' o 'difícil' obtenido será más robusto, en cuanto que vendrá representado por sistemas de recuperación que presentan enfoques distintos.

Algoritmo de las k-medias

Una aproximación sencilla pero que cumple los requisitos planteados a lo largo de esta sección, sería la aplicación del algoritmo de las k-medias (MacQueen (1967), pág 281-297) para realizar la agrupación de consultas.

El algoritmo de las k-medias es un algoritmo básico de agrupamiento, cuyo principal objetivo es construir un número de k particiones, que deben ser indicadas a priori. De esta forma se minimizan las distancias entre los elementos dentro de una misma partición, es decir la distancia a la media de dicha partición entre los elementos es mínima, mientras se maximiza la distancia entre las medias de las particiones.

Dado un conjunto de n consultas y sus valores de precisión media $(t_1, t_2, \cdots, t_n)$, el algoritmo de las k-medias dividirá el conjunto de consultas en k particiones C_i tal que $(k < n)$, $C = \{C_1, C_2, \ldots, C_k\}$, de forma que la distancia intra-partición quede minimizada.

$$\arg\max_{C} \sum_{i=1}^{k} \sum_{t \in C_i} (t - \bar{C_i})^2$$

donde $\bar{C_i}$ es la media en C_i.

El algoritmo de las k-medias tiene como argumentos el número deseado de particiones k en las que se dividirán el conjunto de datos y el valor de las medias de las k particiones que se toman como valores por defecto para inicializar el algoritmo.

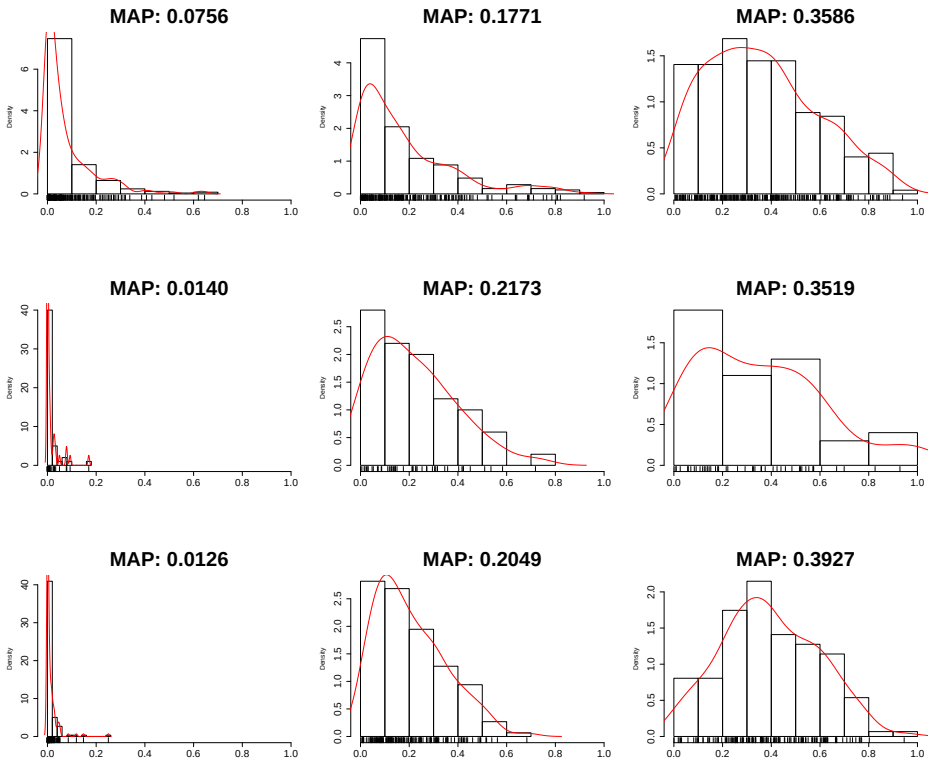

Figura 4.2: Histograma de precisión media y función de densidad para los peores, promedios y mejores sistemas que participaron en la tarea *Robust2004* y en las colecciones *WT10g* y *GOV2*. Cada tarea/colección corresponde a una fila y cada columna representa un grado de calidad del sistema.

Una de las características principales del algoritmo de las k-medias es que es un método no determinista, ya que su resultado depende en gran medida de la inicialización de las particiones.

Para evitar esta indefinición en el resultado de las particiones que se obtengan, se propone fijar el conjunto de medias de las particiones iniciales según la siguiente ecuación:

$$\bar{C}_i = percentil\left(100 \cdot \frac{i}{n+1}\right) \tag{4.1}$$

donde $\bar{C}_i$ es la media inicial de la *i-ésima* partición, y n es el número de particiones deseado. Por ejemplo, para el caso $n = 3$, las medias de inicialización del algoritmo serán fijadas al *25*, *50* y *75* percentil de los datos.

El objetivo de fijar las medias de inicialización según la Ecuación 4.1, es evitar que se introduzca algún tipo de sesgo en las construcción de las particiones. Así, éstas se ajustarán en lo posible a la distribución de valores presente en el conjunto de datos a dividir.

La Ecuación 4.1 distribuye de manera uniforme el conjunto de medias de inicialización a lo largo del conjunto de consultas, de forma que el número de consultas entre las distintas medias sea similar.

El uso del algoritmo de las k-medias, junto con la aproximación sugerida para la inicialización del algoritmo, asegura que los grupos se construirán de la forma más fiel posible a la distribución de los datos.

Este hecho puede ser observado en la Figura 4.3, donde queda reflejado como aquellas particiones que contienen un mayor número de consultas, corresponden a aquellas zonas donde la función de densidad de los datos muestra un máximo, coincidiendo éstas con el conjunto de consultas con menor valor de precisión media.

Se considera que el número de particiones entre las cuales se dividirán las consultas debe ser un parámetro libre, que depende del grado de granularidad deseado a la hora de realizar el estudio [3]. Un escenario típico podría incluir tres grupos de consultas 'Fáciles', 'Medias' y 'Difíciles'. Una restricción a este modelo, es que para poder comparar resultados entre distintos métodos de predicción, el número de particiones, así como los valores de calidad reales de las consultas, deben ser fijados de forma equivalente independientemente del método que se desea evaluar.

Otra característica importante de este nuevo marco de evaluación es que para que un método de predicción obtenga resultados óptimos, la distribución de las predicciones deberá ser lo más similar posible a la distribución que tengan los valores de precisión media.

Esta propiedad es consecuencia de que el cálculo del porcentaje de aciertos, depende en cierta medida del número de consultas por partición. Así,

[3] Aunque existen diversas aproximaciones para fijar dicho valor de forma automática.

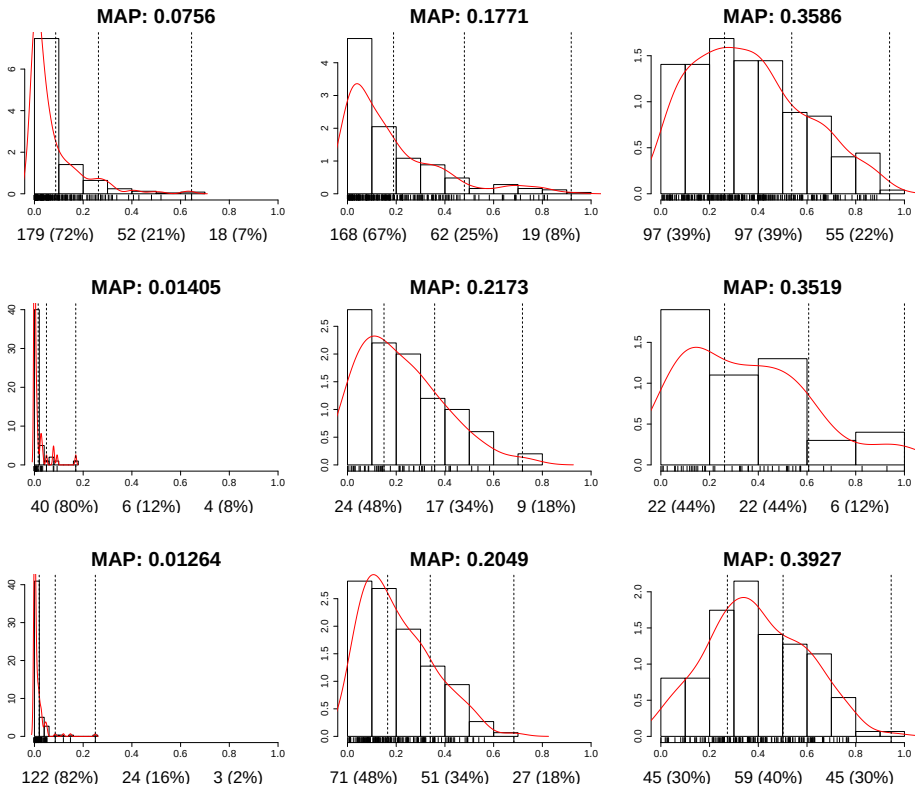

Figura 4.3: Histograma de precisión media y función de densidad para los peores, promedios y mejores sistemas que participaron en la tarea *Robust2004* y en las colecciones *WT10g* y *GOV2*. Cada tarea/colección corresponde a una fila y cada columna representa un grado de calidad del sistema. En este caso aparece además el resultado de aplicar el algoritmo de las k-medias siendo $k = 3$. Cada histograma aparece dividido en tres particiones, junto con el número de elementos y porcentaje sobre el total de consultas que han sido asignadas a cada partición.

la probabilidad de acierto se incrementa cuando las particiones equivalentes en términos de calidad según la predicción y la precisión media, tienen un número similar de elementos. Como se ha descrito en detalle con anterioridad el número de elementos en una partición es consecuencia directa de la distribución de los datos utilizada.

En la siguiente sección se aplica el nuevo marco de evaluación a un subconjunto de diferentes métodos de predicción en una colección estándar de *TREC*, con el objetivo de validar la propuesta presentada y su habilidad para evitar algunas de las limitaciones que los métodos de evaluación actuales presentan.

4.4. Experimentos y resultados

Para comprobar la utilidad del método de evaluación presentado, se opta por utilizar el conjunto de 249 consultas utilizadas en la tarea *Robust2004* (Voorhees (2004)). Como ya se mencionó en el capítulo anterior, se considera que esta colección y la tarea *Robust2004* en particular, es la más adecuada, ya que es el conjunto de consultas donde los distintos métodos de predicción muestran un mayor grado de acierto.

Se presentan dos experimentos basados en los dos escenarios descritos anteriormente. En el primero se realiza la evaluación respecto a un sistema de búsqueda concreto, que corresponderá con los resultados devueltos utilizando la función de ranking *Query Likelihood*.

A continuación, se repite el mismo experimento, pero en este caso utilizando un sistema promedio cuyos valores de precisión media se obtienen calculando el valor medio obtenido por los 110 sistemas que participaron en dicha competición en el año 2004. Este valor promedio puede ser interpretado como aquel valor que representa de forma más fidedigna la dificultad o calidad de una consulta específica, ya que dicha dificultad viene expresada por los valores obtenidos por un conjunto muy amplio de sistemas que aplican enfoques de recuperación muy diversos. En este caso, al no disponer de un conjunto de documentos devueltos por consulta que represente el total de los sistemas utilizados, se opta por utilizar como lista de resultados la misma que en el escenario anterior. Debe tenerse en cuenta que dicho condicionante, puede afectar en cierta medida al rendimiento de los métodos *Post-Retrieval*, por el uso que este tipo de métodos realizan de dicha lista.

A continuación se implementaron un conjunto de métodos de predicción para comparar su rendimiento en base al marco de evaluación propuesto. Para el caso *Pre-Retrieval* se optó por incluir en la evaluación aquellos propuestos por He et. al (He y Ounis (2004)). Estos se refieren principalmente a aproximaciones basadas en los valores de *IDF* o *ICTF* de los términos de la consulta. Así, se aplican: Máximo *IDF*, *ICTF* promedio, *Simplified Clarity Score (SCS)* y *QueryScope* basado en el número de documentos recuperados

por cada uno de los términos de la consulta.

Para el caso de los *Post-Retrieval* se opta por utilizar *Clarity Score* (Cronen-Townsend et al. (2002)), fijando el número de documentos para construir el modelo de lenguaje de la consulta en 500, y el método presentado en el capítulo anterior basado en la desviación de los documentos.

El número de documentos sobre los que se calcula la desviación, se obtiene utilizando como estimador la función de escalado sobre el operador *AND*, con un factor de escala $\lambda = 5$. Como configuración general se opta por establecer tres tipos distintos de consultas: 'Fáciles', para aquellas consultas que obtengan valores altos de precisión media; 'Difíciles', para las consultas con un valor bajo de precisión media; y finalmente 'Medio': para el resto de consultas que no hayan sido incluidas en alguno de los grupos anteriores.

Las particiones resultado de aplicar el algoritmo de las k-medias al conjunto de consultas quedan reflejadas en la Tabla 4.2 para el caso del único sistema y 4.3 para el caso del conjunto de sistemas.

En ambos casos y como era de esperar, por las distribuciones de datos que conforman los valores de precisión media obtenidos, el grupo mayor de consultas corresponde al de las consultas 'Difíciles', mientras que el menor grupo es el de las consultas consideradas 'Fáciles'.

Tabla 4.2: Se incluye el número de consultas que conforma cada grupo, y los valores máximo, mínimo, media y desviación estándar de la precisión media en cada uno de los grupos.

	Núm	**Máx**	**Media**	**Mín**	**Desv**
Difícil	121	.19	.07	.0005	.005
Medio	97	.48	.31	.20	.008
Fácil	31	.91	.66	.49	.011

Medidas de evaluación

Las herramientas básicas que permiten observar el rendimiento de los distintos métodos de predicción en distintos grupos de consultas, son las diferentes medidas de evaluación que se apliquen.

Tabla 4.3: Se incluye el número de consultas que conforma cada grupo, y los valores máximo, mínimo, media y desviación estándar de la precisión media en cada uno de los grupos.

	Núm	**Máx**	**Media**	**Mín**	**Desv**
Difícil	111	.20	.11	.007	.05
Medio	91	.42	.30	.20	.05
Fácil	47	.81	.55	.42	.11

En el ámbito de la clasificación existe un amplio rango de medidas disponibles que permiten observar distintos detalles del rendimiento mostrado por un clasificador específico. Así, una primera aproximación consistiría en medir el número, o ratio, de aciertos global y por grupo de consultas. Además, se emplea la clásica medida-F:

$$2 \cdot \frac{precision \cdot cobertura}{precision + cobertura}$$

Esta medida puede ser aplicada por grupos o de manera global para el total de las consultas.

Distance Based Error Measure (DBEM)

Una característica común que comparten el conjunto de medidas previas es la forma en la que tratan los errores. Así, para estas medidas cada error en la clasificación supone la misma penalización, ya que no hacen distinción entre algunas diferencias evidentes que aparecen con los elementos que se clasifican de manera incorrecta.

El problema principal es que este tipo de medidas ignoran el concepto de distancia entre grupos. Supongamos una configuración en la cuál tenemos tres tipos de particiones: 'Fácil', 'Medio' y 'Difícil'. Predecir que una consulta es 'Fácil' cuando realmente es 'Difícil', debería tener una penalización mayor que en el caso que hubiera sido predicha como 'Media', ya que la partición correcta 'Difícil' es más cercana a la partición 'Media' que a la partición 'Fácil'.

Para resolver este tipo de problemas en la evaluación de los métodos de predicción, se propone una nueva medida que incluye la distancia entre grupos. Esta nueva medida se denomina *Distance Based Error Measure (DBEM)* y nos proporciona una medida global del rendimiento de un método de predicción según diferentes tipos de consultas, pero centrada en el caso de elementos clasificados de forma incorrecta. Formalmente, dado un conjunto de consultas dividido en k grupos, se define la distancia entre dos de esos grupos de consultas C_i y C_j como:

$$distancia(G_i, G_j) = \|i - j\|$$

donde $0 < i,j \leq k$. Entonces:

$$DBEM = \frac{\sum_i^n distancia(P_{t_i}, C_{t_i})}{\sum_i^n max\left[\forall_{j \in n} distancia(P_{t_i}, P_{t_j})\right]}$$

donde P_{t_i} es la partición a la que pertenece la consulta t_i según el método de predicción, C_{t_i} es la partición a la que pertenece la consulta t_i según la precisión media y n es el número total de consultas. Es decir, el valor obtenido con la medida *DBEM*, es equivalente a la distancia entre todas las consultas normalizada por la máxima distancia posible entre todas las consultas,

donde distancias menores implican un nivel más elevado de rendimiento del método de predicción.

Resultados

En la Tabla 4.4, aparecen los resultados de aquellas consultas correctamente clasificadas junto con el número de consultas en la partición.

Los resultados indican que el conjunto de métodos de predicción basados en *IDF*, tienen un fuerte sesgo que hace que tiendan a agrupar la mayoría de las consultas como 'Difíciles'.

Esto les permite, obtener los mejores resultados a la hora de predecir consultas 'Difíciles', aunque esta misma razón provoca que los resultados de las consultas consideradas como 'Medias' o 'Fáciles' sean muy pobres. Además se puede observar que dicha familia de métodos de predicción obtienen resultados muy similares entre sí, como aparece al analizar el resultado que obtienen para todo el conjunto de consultas (0,48). La similitud en los resultados de este tipo de métodos no es capturada por los coeficientes de correlación. Es más, los resultados que obtienen en términos de correlación, podrían conducir a conclusiones equivocadas o contrarias, ya que el grado de correlación que muestran es muy diverso. Por ejemplo, de acuerdo al valor de correlación *Pearson* obtenido por *IDFMin* ($r = 0{,}24$), este método mejora claramente a *IDFAvg* ($r = 0{,}17$). Sin embargo, con los valores que obtienen de acuerdo a *Kendall* la conclusión debería ser exactamente la contraria, ya que *IDFMin* obtiene un valor τ de 0.16 frente al $\tau = 0{,}32$ que alcanza *IDFAvg*. Sin embargo, esta conclusión quedaría descartada basándonos en el total de aciertos que obtienen cualquiera de los métodos basados en *IDF*, ya que son prácticamente equivalentes, alrededor de 122 aciertos.

De esta forma se pueden extraer conclusiones más precisas respecto al rendimiento de dichos métodos. Por ejemplo, la precisión global obtenida por el método *QueryScope* indica que este método presenta un rendimiento inferior al de aquellos basados en *IDF*. Pero al observar los resultados parciales, se demuestra que dicho método obtiene un mejor rendimiento para las consultas de tipo 'Fácil' y 'Medio' en comparación con los métodos basados en *IDF*.

Con la observación de los resultados, se puede concluir que el gran rendimiento que presentan los métodos basados en *IDF* es simplemente la consecuencia de que consideran a más del 95 % de consultas como 'Difíciles'. Esto les permite conseguir un alto grado de cobertura en este tipo de consultas, cercano al 100 %, pero con una precisión muy baja ya que más del 50 % de consultas que son predichas como 'Difíciles' en realidad no lo son. Claramente este sesgo no refleja la realidad, ya que el número de consultas consideradas como 'Difíciles' según la precisión media, está por debajo del 50 % del total de consultas en ambos escenarios.

Finalmente, los resultados basados en el número de aciertos confirman

que los métodos que muestran un mayor rendimiento, como era de esperar, corresponden a *Clarity Score* y al método basado en la desviación estándar (*ScoreDesv*), mostrando ambos un rendimiento muy superior al resto.

Tabla 4.4: Las primeras tres columnas muestran el número de aciertos por partición, incluyendo entre paréntesis el número total de consultas etiquetadas con el tipo de la columna en la que aparecen. La cuarta columna representa el total de aciertos. Finalmente las tres últimas columnas muestran la precisión total medida como $\frac{aciertos}{total}$ y los valores de correlación *Pearson* y *Kendall* que se han obtenido.

	Difícil	**Medio**	**Fácil**	**Total**	**Prec**	**Pearson**	**Kendall**
AVICTF	73(122)	47(110)	9(17)	129	0.52	0.45	0.26
IDFAvg	120(245)	1(3)	1(1)	122	0.49	0.17	0.32
IDFDesv	119(245)	1(3)	1(1)	121	0.48	0.12	0.25
IDFMax	118(243)	1(5)	1(1)	120	0.48	0.15	0.32
IDFMin	121(245)	1(3)	1(1)	123	0.49	0.24	0.16
QScope	91(171)	22(67)	5(11)	118	0.47	0.37	0.18
SCS	60(87)	50(109)	18(53)	128	0.51	0.45	0.26
CS	78(110)	54(103)	13(36)	145	0.58	0.51	0.41
ScoreDesv	74(104)	59(108)	19(37)	152	0.61	0.56	0.38

Como se ha podido observar, la utilización del número de aciertos global como medida de evaluación, puede conducir a extraer conclusiones poco acertadas al igual que ocurre con el uso de coeficientes de correlación.

Este hecho puede ser evitado con el uso de una medida de evaluación más sofisticada, como la medida-F, sobre el conjunto de las consultas.

Los resultados con la medida-F aparecen en la Tabla 4.5. Esta medida destaca de forma clara que los métodos de predicción basados en *IDF*, muestran el rendimiento más pobre para el conjunto de todas las consultas.

Como se puede observar, esta familia de métodos apenas supera un valor de medida-F de 0.33 para el global de las consultas, lo que se acerca al grado de acierto que obtendría un método aleatorio. El siguiente método que muestra un valor más bajo de medida-F, para todas las consultas, es (*QueryScope*) con un valor de 0.43. El estudio de los resultados de medida-F sugiere nuevas conclusiones. Los resultados muestran que en general, los métodos de predicción evaluados detectan más fácilmente las consultas 'Difíciles' respecto a los otros tipos.

Otra conclusión interesante aparece en relación a los métodos *SCS* y *AVICTF*. Ambos obtenían resultados equivalente al ser evaluados con coeficientes de correlación ($r = 0{,}45$ y $\tau = 0{,}26$). Sin embargo, en cuanto a la medida-F se observa que *SCS* es capaz de detectar hasta un 16 % más de consultas 'Fáciles' que *AVICTF*.

También aparecen ciertas diferencias entre los dos métodos que muestran un mejor rendimiento, *CS* y *ScoreDesv*. Así, usando la medida-F por grupos se puede comprobar que mientras *CS* es ligeramente superior a *ScoreDesv*

en relación a la detección de consultas de tipo 'Difícil', este último método mejora a *CS* en cuanto a la predicción de consultas de tipo 'Fácil', ya que incrementa la detección de este tipo de consultas hasta en un 43 %. Por tanto, aparentemente *ScoreDesv* es un método más adecuado para aquellas situaciones en las que el interés se centrara especialmente en la detección de consultas con un elevado valor de precisión media.

Tabla 4.5: Las primeras tres columnas muestran el valor de medida-F para cada partición, las últimas cuatro columnas muestran la medida-F para el conjunto de consultas, la *'Distance Based Error Measure'* y los valores de *Pearson* y *Kendall* que se han obtenido.

	Difícil	**Medio**	**Fácil**	**Total**	**DBEM**	**Pearson**	**Kendall**
AVICTF	0.60	0.45	0.37	0.51	0.32	0.45	0.26
IDFAvg	0.65	0.02	0.06	0.33	0.39	0.17	0.32
IDFDesv	0.65	0.02	0.06	0.33	0.39	0.12	0.25
IDFMAX	0.63	0.02	0.06	0.33	0.39	0.15	0.32
IDFMin	0.66	0.02	0.06	0.36	0.38	0.24	0.16
QScope	0.62	0.26	0.23	0.43	0.35	0.37	0.18
SCS	0.58	0.48	0.43	0.52	0.34	-0.45	-0.26
CS	0.67	0.54	0.39	0.59	0.29	0.51	0.41
ScoreDesv	0.65	0.57	0.55	0.61	0.27	0.56	0.38

Pequeñas diferencias de rendimiento entre *CS* y *ScoreDesv* también se reflejan con el uso de la medida propuesta *DBEM*. Los resultados obtenidos sugieren un menor ratio de fallo para *ScoreDesv*, hecho este que puede ser confirmado al observar la matriz de confusión para ambos métodos que se muestra en la Tabla 4.6. En esta tabla se puede ver que mientras el número de consultas incorrectamente etiquetadas es 104 para *CS* y 97 para *ScoreDesv*, *CS* predice 12 consultas 'Difíles' como 'Fáciles' o viceversa, mientras este tipo de error ocurre 10 veces con la utilización de *ScoreDesv*. Es decir, este ratio de error implica que la proporción de errores de etiquetado de la 'máxima distancia' es alrededor de un 11 % para *CS* y de un 10 % para *ScoreDesv*.

Tabla 4.6: Matriz de confusión para *Clarity Score* (izquierda) y *ScoreDesv* (derecha), el número de consultas correctamente predichas aparece en negrita.

	Clarity Score					**ScoresDesv**			
	Difícil	**Medio**	**Fácil**	**Total**		**Difícil**	**Medio**	**Fácil**	**Total**
Difícil	**78**	36	7	121	**Difícil**	**74**	39	8	121
Medio	27	**54**	16	97	**Medio**	28	**59**	10	97
Fácil	5	13	**13**	31	**Fácil**	2	10	**19**	31
Total	110	103	36	**145**	**Total**	104	108	37	**152**

Finalmente, debe destacarse el grado de correlación encontrado entre las

nuevas medidas de evaluación que actúan a nivel global y las medidas de correlación utilizadas actualmente para la evaluación de métodos de predicción, como se recoge en la Tabla 4.7.

Esta propiedad del nuevo marco de evaluación es de gran importancia ya que pone de manifiesto que el nuevo marco propuesto recoge también la información que aportan los coeficientes de correlación.

Por lo tanto este nuevo marco de evaluación no solo es capaz de suministrar una información equivalente a aquella que se obtiene mediante el uso de coeficientes de correlación, sino que además es capaz de extraer una información más detallada respecto al comportamiento de los métodos de predicción frente a distintos tipos de consultas, siendo esta una de las principales motivaciones para el desarrollo de este nuevo marco de evaluación.

Tabla 4.7: Grado de correlación *Pearson*, entre las medidas globales de evaluación utilizadas y el enfoque clásico basado en coeficientes de correlación.

	Precisión	**Medida-F**	**DBEM**
Pearson	0.78	0.77	-0.95
Kendall	0.77	0.66	-0.57

Escenario Sistema Genérico

Los resultados obtenidos con el segundo escenario propuesto, es decir aquel en el que la calidad de una consulta viene dada como el promedio de la precisión media obtenida a partir de un conjunto amplio de sistemas, aparecen en la Tabla 4.8 y en la Tabla 4.9.

Los resultados que se obtienen en este segundo escenario son muy similares a los obtenidos anteriormente para el caso de los métodos *Pre-Retrieval*. Sin embargo, los métodos *Post-Retrieval*, es decir *Clarity Score* y *ScoreDesv*, muestran un leve descenso de rendimiento. Esta perdida de rendimiento, es consecuencia del hecho de que ambos métodos utilizan la lista de resultados para calcular sus predicciones y en este escenario no se dispone de una lista de resultados que represente al conjunto de los sistemas.

Es más, *ScoreDesv* presenta un descenso de rendimiento más acusado debido a su propia naturaleza e importante dependencia de la lista de resultados. Sin embargo, *Clarity Score* utiliza dicha lista de resultados para calcular un modelo de lenguaje de la consulta, por lo que se ve menos afectado al utilizar una lista de resultados que no corresponde exactamente con los valores de precisión media obtenidos.

Esta misma razón explica el hecho de que los métodos *Pre-Retrieval* no vean disminuido su rendimiento, ya que no dependen de la lista de resultados. Al contrario, algunos de los métodos *Pre-Retrieval*, como *SCS* y *AVICTF*, aumentan de forma muy leve su rendimiento.

Tabla 4.8: Las primeras tres columnas muestran el número de aciertos por partición, incluyendo entre paréntesis el número total de consultas etiquetadas con el tipo de la columna en la que aparecen. La cuarta columna representa el total de aciertos. Finalmente las tres últimas columnas muestran la precisión total medida como $\frac{aciertos}{total}$ y los valores de *Pearson* y *Kendall* que se han obtenido.

	Difícil	**Medio**	**Fácil**	**Total**	**Precisión**	**Pearson**	**Kendall**
AVICTF	72(122)	49(110)	11(17)	132	0.53	0.47	0.25
IDFAvg	111(245)	1(3)	1(1)	113	0.45	0.13	0.30
IDFDesv	110(245)	1(3)	1(1)	112	0.45	0.11	0.25
IDFMax	109(243)	1(5)	1(1)	111	0.45	0.12	0.30
IDFMin	111(245)	2(3)	1(1)	114	0.46	0.10	0.13
QScope	87(171)	25(67)	5(11)	117	0.47	0.28	0.17
SCS	59(87)	53(109)	22(53)	134	0.54	0.38	0.26
CS	73(110)	47(103)	19(36)	139	0.56	0.52	0.40
ScoreDesv	68(104)	50(108)	21(37)	139	0.55	0.55	0.36

Tabla 4.9: Las primeras tres columnas muestran el valor de medida-F para cada partición, las últimas cuatro columnas muestran la medida-F para el conjunto de consultas, la *Distance Based Error Measure* y los valores de *Pearson* y *Kendall* que se han obtenido.

	Difícil	**Medio**	**Fácil**	**Total**	**DBEM**	**Pearson**	**Kendall**
AVICTF	0.62	0.49	0.34	0.48	0.32	0.47	0.25
IDFAvg	0.62	0.02	0.04	0.22	0.45	0.13	0.30
IDFDesv	0.62	0.02	0.04	0.23	0.45	0.11	0.25
IDFMAX	0.61	0.02	0.04	0.23	0.45	0.12	0.30
IDFMin	0.62	0.02	0.04	0.23	0.45	0.10	0.13
QScope	0.62	0.32	0.17	0.37	0.39	0.28	0.17
SCS	0.60	0.53	0.44	0.52	0.34	0.38	0.26
CS	0.66	0.48	0.46	0.53	0.30	0.52	0.40
ScoreDesv	0.63	0.50	0.50	0.54	0.30	0.55	0.36

4.5. Conclusiones

En este capítulo se ha introducido un nuevo marco de evaluación específico para los métodos de predicción del rendimiento de consultas. El principal objetivo de dicho marco era evitar las limitaciones que aparecen en una evaluación basada en coeficientes de correlación clásicos como *Pearson* o *Kendall*.

Este tipo de limitaciones se pueden resumir en los siguientes puntos:

- *Pearson* posee ciertas características que no lo hacen adecuado a la hora de evaluar el rendimiento de métodos de predicción, tales como el efecto que producen la presencia de datos atípicos o la existencia de relaciones no lineales entre las series de datos.

- La evaluación basada en correlación considera equivalentes a todas las consultas, ya que cada una de estas consultas supone la misma aportación a la medida final de correlación, lo que en algunos casos puede no ser aconsejable.

- Las aproximaciones basadas en coeficientes de correlación producen resultados que en muchos casos son muy similares para distintos métodos de predicción y por tanto difíciles de interpretar.

- Este tipo de evaluación no está enfocada a la aplicación de los métodos de predicción a un contexto específico, ya que los resultados que se obtienen no son específicos de aquellas consultas o tipo de consultas de interés para su aplicación.

- Se puede concluir que en algunos casos los coeficientes de correlación dan una visión poco detallada del rendimiento de un método de predicción.

Con la propuesta presentada en este capítulo se evitan las limitaciones descritas, mediante la transformación del problema de evaluación de métodos de predicción en un problema de clasificación. Esto se consigue asumiendo que cada una de las consultas pertenece de forma unívoca a un tipo específico de consultas, dependiendo del rendimiento mostrado por dicha consulta.

Para realizar esta labor se ha propuesto un método automático de agrupación de consultas basado en la calidad de éstas sobre un doble escenario: de acuerdo al rendimiento mostrado por un sistema de recuperación de información específico y de acuerdo a la calidad de una consulta dada por un conjunto de sistemas de recuperación distintos. Es decir, un enfoque que depende de un sistema único y un enfoque en el que se intenta cuantificar la calidad de las consultas en base a las respuestas obtenidas a partir de sistemas diversos.

Mientras que con la aplicación de los coeficientes de correlación se ha observado que se ocultan ciertos detalles del rendimiento de los métodos de predicción, la nueva aproximación hace explícitas estas diferencias suministrando información que puede ayudar a la selección de un método u otro dependiendo del contexto donde se desee aplicar. Además, ya que cada consulta es etiquetada a partir de la predicción, en base a esta etiqueta un sistema podría decidir aplicar de forma automática diferentes técnicas para mejorar la calidad de la respuesta. Sin embargo, con el uso de coeficientes de correlación, esta decisión debería ser tomada en base al valor numérico asignado por el método de predicción en uso.

La propuesta presentada en este capítulo ha sido evaluada contra un conjunto de diferentes métodos de predicción, lo que ha permitido extraer conclusiones sobre el rendimiento de dichos métodos de predicción dependiendo del tipo de consultas a las que intenten predecir.

Otra de las características principales del marco de evaluación propuesto se refiere al extenso número de medidas de evaluación que pueden ser aplicadas. Así, al utilizar el símil de un problema de clasificación, se hace posible la utilización de medidas utilizadas habitualmente en este campo que son de sobra conocidas y cuya conveniencia y robustez han sido probadas.

Como extensión al uso de las medidas clásicas, se propone una nueva medida (*DBEM*) que está orientada de forma específica al tipo de errores en los que se incurre al predecir de forma errónea una consulta. De esta forma, es posible considerar distintos grados en los errores y así estimar qué método de predicción es más adecuado para un contexto específico.

Finalmente debe destacarse que las medidas globales propuestas dentro del marco de evaluación alcanzan un alto grado de correlación con los métodos clásicos de evaluación (*Pearson* y *Kendall*). Por tanto, el marco de evaluación propuesto es capaz de suministrar una información similar a la que suministran los coeficientes de correlación y simultáneamente mostrar de forma más detallada el rendimiento de estos en un contexto específico.

Never question the relevance of truth, but always question the truth of relevance.

Craig Bruce.

Capítulo 5

Expansión selectiva de consultas

5.1. Introducción

El propósito de este capítulo es realizar un estudio del rendimiento que se obtiene con el uso de métodos de predicción en el campo de la expansión de consultas sin información de usuario.

Este tipo de expansión, que se suele denominar *"pseudo relevance feedback" (PRF)*, ha sido utilizada ampliamente en la comunidad de recuperación de información con el objetivo de mejorar la calidad de la respuesta dada a un usuario (Xu y Croft (1996); Evans y Lefferts (1993)). Este objetivo se consigue refinando la consulta original expresada por el usuario, lo que generalmente implica añadir ciertos términos a la consulta original de manera automática, con la intención de recuperar un mayor número de documentos relevantes.

La expansión de consultas en ausencia de información por parte del usuario es probablemente el caso de uso más referenciado en el campo de la predicción de la calidad de las consultas. Este caso de uso se menciona de forma habitual como el escenario típico donde el uso de las estimaciones obtenidas a partir de un método de predicción permitiría mejorar el rendimiento de los sistemas de recuperación. La mejora sería consecuencia de la posibilidad de estimar con cierto grado de precisión, en qué casos al expandir una consulta el conjunto de documentos recuperados será más relevante que con el uso de la consulta original.

El incremento de rendimiento que se espera con la integración de los métodos de predicción y las técnicas clásicas de expansión de consultas se debe a una limitación inherente a los métodos actuales de expansión. Esta limitación aparece como consecuencia de la degradación de los resultados en algunas consultas al ser expandidas. Así, mientras para algunas consultas la expansión produce un importante incremento en la calidad de las respuestas

que se obtienen, para otro conjunto, no menos importante, la calidad de los resultados disminuye de manera significativa. Idealmente se espera que con la aplicación de los métodos de predicción se pueda prever en qué casos una consulta expandida producirá un rendimiento menor al de la consulta original y por tanto tener la capacidad de evitar la expansión de dicha consulta.

De esta forma se podría evolucionar desde la aplicación típica de los métodos de expansión que se realiza de forma sistemática a todo el conjunto de consultas, a la denominada expansión selectiva de consultas, donde solo se expanden un subconjunto específico de consultas, y éstas son seleccionadas en base a la información que se obtiene con el uso de los métodos de predicción.

5.2. Expansión de consultas sin información de usuario

La expansión de consultas en ausencia de información de usuario o *PRF* es un tipo de retroalimentación de la consulta en la cuál no es necesaria la intervención del usuario, ya que la selección de los denominados términos candidatos, que servirán para realizar la expansión de la consulta original Q, se realiza de manera automática. En este tipo de expansión la selección de los términos se obtiene a partir de los primeros k documentos que hayan sido devueltos basándose en la consulta original. La intuición detrás de este modelo es asumir que dicho conjunto de k documentos recuperados son relevantes, y por tanto algunos de los términos que aparecen en los documentos que pertenecen a k representarán de forma fidedigna el asunto expresado por el usuario en la consulta original Q. Con la inclusión en la consulta original de este subconjunto de términos se pretende incrementar la capacidad de la consulta de recuperar nuevos documentos relevantes. Así, partiendo de la consulta original Q se construye una consulta expandida Q' que contiene algunos términos extra que aparecen en k. Este proceso puede ser dividido en dos etapas principales:

- Seleccionar un conjunto de términos candidatos para la expansión del conjunto de documentos k.
- Asignar un grado de representatividad o peso numérico a cada uno de los términos candidatos. Dicho peso se asigna en función del grado de 'importancia' de cada término en relación a la temática específica de la consulta.

Asignar un peso a cada uno de los términos de la consulta restringe el efecto que cada uno de los nuevos términos añadidos a la consulta original tiene en el conjunto de documentos recuperados. Este paso es de gran

importancia ya que la inclusión de términos nuevos en la consulta puede desviar, (*query drift*), de forma importante el objetivo original que el usuario intentó expresar en su consulta en lugar de refinarla. Así, asignar un peso a los términos permite evitar este efecto, ya que en general a los términos de expansión se les suele asignar un grado de importancia (peso) considerablemente menor que a los términos originales, respetando el espíritu original de la consulta expresada por el usuario.

Existen un gran número de alternativas propuestas para realizar tanto la selección de términos como el denominado repesado de estos. Uno de los métodos clásicos que ha mostrado un mejor rendimiento es el basado en la divergencia *Kullback-Leibler* (Carpineto et al. (2001), Parapar y Barreiro (2011)). Este método ha sido aplicado asiduamente en competiciones como *TREC* (Lioma et al. (2006)) con un éxito considerable.

Expansión de consultas basada en la divergencia *(KL)*

El objetivo principal que se pretende alcanzar con el uso de la divergencia *Kullback-Leibler* es extraer los términos más discriminantes de los primeros k documentos devueltos con la consulta original. Un término se considera muy discriminante en el caso de que aparezca con mucha frecuencia en los primeros k documentos mientras que su presencia en el total de la colección sea poco significativa. Es decir, aquellos términos considerados discriminantes serán los términos específicos del lenguaje utilizado en el conjunto de los k documentos devueltos en primer lugar.

De forma general *KLD* calcula el grado de divergencia entre dos distribuciones de probabilidad. Así, utilizando *KLD* se estima la divergencia entre la probabilidad de que un término t aparezca en el subconjunto de documentos k (p_k) y la probabilidad de que dicho término aparezca en la colección de documentos total C (p_C). La ecuación queda de la forma:

$$\mathrm{KLD}_{p_k,p_C}(t) = p_k(t) \log \left(\frac{p_k(t)}{p_C(t)} \right) \tag{5.1}$$

donde $p_K(t)$ es la probabilidad de que el término t aparezca dentro de los k primeros documentos, calculada como el número de veces que aparece el término t en el conjunto de documentos k que han sido recuperados con la consulta original. Finalmente $p_C(t)$, es la frecuencia del mismo término t, pero esta vez en el conjunto de la colección C.

Con la aplicación de la expresión anterior el conjunto de términos que aparecen en k podrán ser ordenados según el grado de divergencia que muestren respecto al conjunto total de documentos. Así, estableciendo un umbral superior podremos seleccionar aquellos términos con una divergencia mayor y por tanto más adecuados para realizar la expansión de la consulta original.

El resultado obtenido con este método se puede observar en la Tabla 5.1, donde aparece el conjunto de términos candidatos seleccionado para la

consulta *"Alzheimer's Drug Treatment"*, que corresponde a la consulta 339 para la tarea *Robust2004*.

Se puede observar que la lista de términos candidatos para dicha consulta contiene términos desde lo muy específico como *"amyloid"* o *"neurotransmitt"* hasta términos más genéricos como *"victim"*, *"known"*, *"cell"*, pero en general el conjunto de términos candidatos en alguno de sus significados pertenecen al dominio de la medicina, estrechamente relacionado con la consulta original.

Una opción muy habitual a la hora de realizar el repesado de los términos de expansión es utilizar el mismo valor obtenido al calcular la divergencia *KL*. Mientras a los términos originales de la consulta se les asigna un valor equivalente a uno, el subconjunto de aquellos términos con mayor valor de divergencia utilizan el valor de *KLD* que en todos los casos es inferior a uno. De esta forma la introducción de nuevos términos en la consulta modifica de forma ligera el conjunto de documentos recuperados originalmente, evitando desvirtuar el interés original que el usuario pretendía expresar en primer término. Añadir los términos de expansión con un peso equivalente al de su valor de *KLD* a los términos originales de la consulta con un peso de 1 equivale a realizar una interpolación entre términos de expansión y términos originales con $\lambda = 0{,}5$ según la siguiente expresión:

$$Q' = \lambda * Term_{orig} + (1 - \lambda) * Terminos_{exp} \tag{5.2}$$

5.3. Expansión selectiva basada en predicciones

Existe un número importante de trabajos donde se analiza, al menos parcialmente, la eficacia de los métodos de predicción cuando estos son aplicados en el escenario de la expansión de consultas en ausencia de información de usuario.

Dentro de estos trabajos destacan, por orden cronológico, los realizados por Cronen-Townsend et al. (2004). En este trabajo las predicciones se basan en la divergencia existente entre el modelo de lenguaje construido sobre los documentos recuperados con la consulta sin expandir y aquellos recuperados con la consulta expandida. En este caso los resultados que se obtienen con la expansión selectiva, muestran una pobre tasa de mejora ya que aproximadamente en la mitad de los casos la expansión selectiva degrada los resultados, respecto a la expansión sistemática. Se observa que en el mejor de los casos la aplicación de la expansión selectiva produce un incremento en *MAP* desde 0.2013 a 0.2115 para el conjunto de consultas de la 301-350 (*TREC 6*), mientras que en el peor de los casos se empeora el rendimiento pasando de un *MAP* de 0.2514 a otro de 0.2212 para las consultas desde la 351 a la 400 (*TREC 7*).

Posteriormente Amati et al. (2004), proponen un método similar de ex-

Tabla 5.1: Lista de los 40 términos que muestran una mayor divergencia extraídos a partir de los 10 documentos recuperados en primer lugar con la consulta *"Alzheimer's Drug Treatment"*. La primera columna corresponde a la raíz (*"stem"*) de los términos y la segunda al valor de *KLD* que obtienen. Como se puede observar aquellos términos con mayor valor de *KLD*, más discriminantes, aparecen al comienzo de la lista.

alzheim	0.31536285305749
dementia	0.1567107039483545
diseas	0.1408059687558589
brain	0.10152882234386738
drug	0.08034370024655754
amyloid	0.0749817928760831
patient	0.06404161093658585
treatment	0.03811911814761063
symptom	0.03493321776394026
memori	0.03263777729752926
protein	0.03180334548985989
degen	0.030845864308654904
disord	0.02924727295203292
research	0.028619411639784716
parkinson	0.02557228570019061
caus	0.022840430358082496
neurotransmitt	0.02268004016699352
sumatriptan	0.022545580042821284
cognex	0.021871270539525027
cell	0.021525551305622708
receptor	0.02135932028956392
hydergin	0.02093388439896785
age	0.01989447962045959
aami	0.019723540084793323
infarct	0.01906392617792763
app	0.018776575588463868
ondansetron	0.018543504292871212
acetylcholin	0.017568833184277568
plaqu	0.016538279020580835
beta	0.01609770439935924
tacrin	0.01566397727276215
clinic	0.014978064634819124
known	0.014320527769786599
depress	0.014309260148836383
lambert	0.014168910386894128
ssris	0.01401376123587397
progress	0.013903531163630214
placebo	0.013757979004208828
pharmaceut	0.013260748581091232
victim	0.012740664240960575

pansión selectiva basado en el denominado *InfoQ*. Este método de predicción combina la longitud de la consulta con la frecuencia en la colección de los términos de dicha consulta para calcular sus estimaciones respecto a la calidad de una consulta. A partir de cierto umbral para el valor computado de *InfoQ* se decide expandir o por contra dejar la consulta como fue expresada originalmente por el usuario. En este trabajo los resultados muestran la aparición de algunos casos donde se observa una muy leve mejoría en los valores de *MAP* al realizar la aplicación de expansión selectiva, siendo el resultado más positivo pasar de 0.2434 a 0.2456. Dichos valores se alcanzan con un umbral de *InfoQ* establecido de forma que se maximice el grado de mejora sobre el conjunto de consultas *TREC 8*, es decir de la 401-500. Debe reseñarse que en este trabajo se optó por utilizar el campo descripción de la consulta y no el título como se realiza de manera habitual.

Otra aproximación corresponde a la realizada por Yom-Tov et al. (2005). Aquí, el método de predicción está basado en un algoritmo de aprendizaje que utiliza el grado de solapamiento entre los resultados que obtienen diferentes consultas derivadas a partir de la consulta original. Para comprobar experimentalmente la efectividad de dicho método de predicción en un entorno de expansión selectiva, el autor plantea dos escenarios:

- En el primer escenario de expansión selectiva se expandirá o no en base al valor que se obtenga a partir del método de predicción. Por tanto solo aquellas consultas en las que la predicción indique una mayor efectividad serán expandidas, ya que se considera que el conjunto de documentos a partir del que se realiza la retroalimentación es de mayor calidad (Carmel et al. (2002)). Para dicho escenario se utiliza el campo descripción de la consulta. De nuevo los resultados que se obtienen son poco clarificadores ya que el grado de mejora es insignificante, pasando de un valor de *MAP* de 0.284 al expandir en todos los casos a un valor de *MAP* de 0.285 utilizando expansión selectiva.

- Al observar que en algunos casos se obtiene una mejor respuesta al utilizar el campo título de la consulta en lugar del campo descripción, se plantea un escenario en el que se seleccionará, en base a las predicciones que se obtengan, el campo título o el campo descripción de forma automática. Los resultados que se obtuvieron con la expansión selectiva fueron de un 0.294 de *MAP*, mientras que al usar como consulta una combinación de título y descripción se obtuvo un valor 0.295 de *MAP*. Por tanto no se observa mejora significativa. En este caso se usó el conjunto de consultas propuestas en la tarea *Robust2004*.

Otro trabajo realizado en este caso sobre las colecciones Web *WT10g* y *GOV2* se debe a He y Ounis (2007a). En este trabajo se proponía un método de expansión de consultas más sofisticado. En este método de expansión la recuperación de los documentos no se basaba simplemente en el contenido

textual de los documentos, sino que hacía uso de los distintos campos *html* en los que el texto aparecía, así como de los enlaces que existían entre los documentos. Además, los términos de expansión podían ser seleccionados a partir del corpus en uso o de corpus externos como la Wikipedia. El análisis de la utilidad de datos de predicción en un entorno de expansión selectiva se basaba en expandir de forma local o haciendo uso de fuentes de documentos externas. Esta decisión se tomaba en base al valor de *AvICTF* que obtenía cada uno de los tipos de expansión. Los resultados que obtuvieron muestran que el mejor caso se daba con la colección *WT10g*, donde lograban un incremento del valor de *MAP* de 0.2191 a 0.2362, correspondiendo 0.2191 al caso de expansión sin fuentes externas. En relación a la colección *GOV2*, los resultados no son tan positivos ya que no se aprecian diferencias en el valor de *MAP* usando o no expansión selectiva, obteniendo en ambos casos un valor de *MAP* de 0.3528.

Finalmente Hauff (2010), presenta un estudio teórico en el que intenta estimar el valor de correlación que debe obtener un método de predicción genérico, para que su uso en el ámbito de la expansión selectiva de consultas signifique una mejora significativa. El marco experimental se construye en base a la suposición de que solo aquellas consultas con valores más elevados de precisión media serán mejoradas al utilizar métodos de expansión automáticos. Así, Hauff plantea un experimento en el que a partir de un conjunto de n consultas se selecciona de forma aleatoria un conjunto de m consultas que serán ordenadas de acuerdo al valor de precisión media que obtengan. El ranking ordenado de m consultas se denominará q_{base}. Para obtener un conjunto de predicciones sobre este ranking, se realiza una serie de permutaciones en el orden de las consultas que pertenecen a q_{base}, de forma que cada nueva permutación corresponderá con el orden propuesto por un método de predicción virtual. Con el conjunto de permutaciones se consigue obtener un rango de predicciones que tomarán valores de correlación τ desde 0 a 1.

Para simular el conjunto de consultas expandidas a partir de q_{base} se crea un ranking expandido que cumpla las siguientes condiciones:

- Se define un conjunto de consultas θ que pertenecen a q_{base} tal que contiene a aquellas consultas que obtienen un mayor valor de precisión media. Para establecer el subconjunto θ se define un umbral inferior de AP, tal que todas aquellas consultas que lo superen pertenecen a θ.

- Para cada consulta perteneciente a θ se busca una nueva consulta q_{exp} en el conjunto de n consultas, tal que obtenga un valor superior de AP respecto a la consulta original en θ. Así, para el conjunto de consultas q_{base} que pertenecen a θ se tiene que $q_{base}^{AP} < q_{exp}^{AP}$.

- Para el resto de consultas del ranking que no pertenecen a θ se asignan consultas q_{exp} tal que $q_{base}^{AP} > q_{exp}^{AP}$.

- Se impone a su vez que el valor de MAP del ranking expandido mejore entre un 15 % y un 30 % al valor de MAP que obtenga el ranking base.
- A su vez se impone que el valor óptimo de expansión, es decir aquel valor de precisión media que se obtiene al expandir solo aquellas consultas en las que la expansión implica incremento de AP, que en este caso coincide con las consultas en θ, no debe superar en más de un 3 % al resultado que se obtiene al expandir todas las consultas de forma sistemática.

Con los datos generados con la metodología descrita es posible estimar a partir de qué valor de correlación la expansión selectiva mejora los resultados de la expansión sistemática. Dicho valor se establece en el entorno de $\tau = 0{,}4$. Como la misma autora resalta esta primera aproximación plantea una limitación clara. Se trata de una consecuencia de asumir que aquellas consultas que obtienen un valor elevado de AP, las consultas que pertenecen a θ, siempre mejoran su precisión media al ser expandidas.

Para superar los problemas de este experimento, Hauff plantea un segundo experimento en el cuál introduce de forma aleatoria cierto grado de perturbación en las consultas expandidas, de forma que cierto porcentaje de las consultas que pertenecen a θ obtienen un valor de precisión media menor a aquel que obtendrían si no fueran expandidas. En este segundo caso Hauff comprueba experimentalmente que el grado de correlación necesario para observar mejoras en el rendimiento de la expansión selectiva sobre la sistemática se sitúa en el entorno de $\tau = 0{,}7$.

Aunque con este segundo experimento Hauff se acerca más a la realidad, ya que no asume que aquellas consultas que superan un umbral en términos de precisión media siempre mejoran al ser expandidas, no ocurre lo mismo en el caso de las consultas bajo dicho umbral de precisión media. Así en este segundo experimento, se continúa asumiendo que aquellas consultas que no superan cierto umbral de precisión media cuando son expandidas ven disminuir su rendimiento. Como se comprobará posteriormente este hecho no ocurre en la realidad ya que aunque habitualmente se ha supuesto que la expansión de consultas con baja precisión media siempre empeora el resultado de las consultas al ser expandidas, este comportamiento no es tan general como pudiera parecer.

Después de realizar un recorrido por los trabajos más destacados en el entorno de la expansión selectiva de consultas basada en predicciones, se puede observar que los resultados son poco concluyentes especialmente debido a las siguientes razones:

- Las consultas utilizadas, bien título o descripción, así como el conjunto de colecciones sobre las que se experimenta son muy diversas y por tanto es complicado extraer algún tipo de conclusión general más allá de la percepción de un rendimiento limitado.

- Algo similar ocurre con el uso de diferentes métodos de expansión en cada uno de los trabajos.
- No existe un análisis que pretenda dar explicación al hecho de la falta de rendimiento de la expansión selectiva basada en métodos de predicción, para el caso de expansión en ausencia de información de usuario.

Por tanto, se plantea la necesidad de realizar un análisis sobre esta falta de rendimiento. Para ello se realiza un análisis basado en colecciones estándar *TREC*, junto con la aplicación de un método de expansión clásico como el basado en la divergencia *Kullback-Leibler* (Kullback y Leibler (1951)).

5.4. Datos experimentales

En la Tabla 5.2 aparecen los resultados que se obtienen al realizar la expansión basada en *KLD*. Con el objetivo de plantear un escenario lo más común posible se fijan los parámetros de expansión a valores considerados estándar. Así, el tamaño del subconjunto de documentos de los que se extraerán los términos corresponde a los 10 primeros documentos. Además, se introducen en la consulta original los cuarenta términos que muestren un mayor valor de divergencia *KL*.

Como se puede observar en la Tabla 5.2, los resultados de realizar la expansión mejoran de forma significativa en todos los casos salvo para *WT10g* donde la expansión reduce el rendimiento levemente. Este rendimiento es el esperado para los casos de *Robust2004* y *GOV2*, ya que en ambas colecciones se observa habitualmente una mejora en los resultados con la aplicación de métodos de expansión locales. Sin embargo, no ocurre lo mismo en la colección *WT10g*, debido en gran parte a ciertas características específicas de ésta. Algunos autores han destacado previamente que la expansión de consultas en *WT10g* tiende a degradar la calidad de los resultados. Este comportamiento se observó en la tarea Web perteneciente al *TREC09* (Hawking (2001)). Posteriormente en algunos trabajos aparecen leves mejoras con la utilización de métodos de expansión no locales (He y Ounis (2007a)), o a través de mecanismos de expansión más sofisticados (Amati et al. (2001); Parapar y Barreiro (2011)).

Para mejorar la comprensión del experimento, en la Tabla 5.2 también aparecen los valores de precisión media que se obtienen aplicando expansión selectiva en el mejor de los casos (límite superior SQE_{Max}) y en el peor (límite inferior SQE_{Min}). El límite superior corresponde con aquel valor que se obtiene aplicando el método de expansión solo en aquellas consultas en las que produce alguna mejora en el rendimiento. De manera inversa, el límite inferior corresponde al resultado obtenido cuando se realiza la expansión solo en aquellas consultas en las que el rendimiento disminuye. Ambos límites

calculados para el conjunto de las tres colecciones, nos permiten observar el hecho de que en el caso de que el método de expansión se muestre efectivo, el valor de precisión media que se obtiene al realizar la expansión para todas las consultas se encuentra más cercano al límite superior que al inferior y que el margen de mejora disponible hasta alcanzar dicho límite superior es reducido. Esto nos lleva a concluir que el método de selección propuesto a la hora de expandir, como aquellos basados en predicción de la calidad de la consulta, deben poseer un grado de efectividad muy elevado para que se observe cierta mejora, en comparación con la expansión sistemática de cada una de las consultas.

Tabla 5.2: Resultados comparativos entre consulta base y consulta expandida, así como los resultados de *SQE* óptimo y peor caso con *SQE*.

	Robust2004	**WT10g**	**GOV2**
Original	0.2412	0.2088	0.2882
Expandida	0.2863	0.2046	0.3078
SQE_{Max}	0.3039	0.2215	0.3224
SQE_{Min}	0.2236	0.1919	0.2736

A continuación en la Tabla 5.3 se muestran los resultados obtenidos con el uso de la expansión selectiva utilizando como criterio la calidad de la consulta. Esta calidad viene definida por el valor real de precisión media que obtiene la consulta, y mediante la etiqueta suministrada por el algoritmo de las k-medias tal como se definió en el capítulo 4. Así, se han establecido tres criterios de expansión:

- Expandir en el caso de que la consulta sea 'Fácil'.
- Expandir en el caso de que la consulta sea 'Media'.
- Expandir en el caso de que la consulta sea 'Difícil'.

Los resultados muestran que en general la expansión selectiva no incrementa el valor de precisión media respecto a la expansión sistemática, salvo en el caso de *WT10g* donde, como ya se observó, la expansión daña el rendimiento general que se obtiene bajo este conjunto de test, y por tanto es de esperar que al reducir el número de consultas que se expanden, el resultado sea levemente superior respecto a expandir el total de consultas. Es más, se prueba como errónea la intuición de que aquellas consultas con valor mayor de precisión media son las más adecuadas para expandir (Carmel et al. (2002)), ya que al expandir las consideradas como *'Fáciles'* se obtiene el valor más bajo de precisión media, mientras que los valores superiores de expansión selectiva se alcanzan cuando se aplica la expansión a aquellas consideradas como *'Difíciles'* o *'Medias'*. Es importante resaltar que este hecho

Tabla 5.3: *MAP* que se obtiene al realizar la expansión selectiva por grupos utilizando como criterio el valor real de precisión media de cada una de las consultas.

	Fáciles	Medias	Difíciles	Todas
Robust2004	0.2432	0.2633	0.2622	0.2863
WT10g	0.2017	0.2104	0.2100	0.2046
GOV2	0.2945	0.2971	0.2926	0.3078

ocurre sin la intervención del método de predicción ya que la clasificación de las consultas se realiza en base al valor real de precisión media.

De forma similar se repite el experimento, pero en este caso utilizando el valor de predicción obtenido con el método basado en la desviación descrito en el Capítulo 3 (*ScoreDesv con* $\lambda = 5$). Los resultados aparecen en la Tabla 5.4. Estos resultados muestran una gran similitud con los de la Tabla 5.3, siendo las pequeñas diferencias que aparecen no significativas. Esto permite concluir que el uso de los valores de calidad de las consultas respecto a las predicciones, que intentan estimar dicha calidad, no supone diferencias significativas. Es decir, en el caso de la expansión selectiva es muy similar utilizar las predicciones a los valores de AP reales.

Tabla 5.4: *MAP* que se obtiene al realizar la expansión selectiva por grupos utilizando como criterio la información del método de predicción *ScoreDesv*.

	Fáciles	Medias	Difíciles
Robust2004	0.2484	0.2609	0.2594
WT10g	0.2061	0.2030	0.2130
GOV2	0.2912	0.2947	0.2983

En ambos casos, la aplicación selectiva produce resultados similares o peores a los obtenidos con la expansión sistemática, poniendo de manifiesto dos hechos no esperados. Por un lado se observa que ni aún con el uso de los valores reales de AP que obtienen las consultas, es posible mejorar los resultados aplicando expansión selectiva respecto a la sistemática. Por otro lado, no resulta más beneficioso realizar la expansión solo para aquellas consultas con un valor superior de precisión media, tal y como se afirma en la literatura relacionada (Carmel et al. (2002)).

En relación a la primera observación, es decir, el hecho de que no se observe una diferencia significativa entre utilizar el valor computado por el método de predicción o usar el valor real que describe la calidad de una consulta, la causa se basa en la falta de relación entre el valor de precisión media de una consulta y el incremento o decremento que produce en este valor su posible expansión.

En realidad si lo que se desea es mejorar la expansión al realizarla de

forma selectiva, el método de predicción debería estar centrado en predecir las diferencias entre realizar o no la expansión de una consulta. Sin embargo, la bondad de un método de predicción se evalúa respecto a la correlación que muestra con la precisión media. Para el caso que nos ocupa se ha comprobado que la correlación existente entre el valor de precisión media antes y después de expandir es $\tau = 0{,}02$ para *Robust2004*, $\tau = -0{,}01$ para *WT10g* y $\tau = 0{,}06$ para *GOV2*. Es decir, no existe correlación. De igual manera los valores de correlación que aparecen entre el valor obtenido por el método de predicción y el incremento o decremento al expandir las consultas es $\tau = 0{,}008$ para *Robust2004*, $\tau = 0{,}09$ para *WT10g* y $\tau = 0{,}03$ para *GOV2*.

La segunda observación de carácter experimental se refiere al hecho de que, al contrario de lo esperado, no parece lo más óptimo expandir aquellas consultas con un mayor valor de precisión media. Este hecho es simplemente la consecuencia de que la expansión de todas las consultas produzca mejor resultado que una expansión selectiva, salvo en el caso de *WT10g*. Puesto que en general, el valor de *MAP* para un conjunto de consultas se incrementa cuantas más de éstas se expandan, es inmediato concluir que lo que producirá mejores resultados será expandir aquel grupo de consultas que contenga mayor número de éstas. Esta es la razón que explica porqué en los experimentos propuestos, los mejores resultados no se obtienen al expandir el grupo *'Fáciles'*, ya que en las tres colecciones este grupo contienen un número menor de consultas.

Para comprobar de forma experimental que realmente el valor de precisión media que obtiene una consulta es poco significativo o irrelevante a la hora de decidir si dicha consulta debe ser expandida, exploraremos el comportamiento de la expansión selectiva en el caso de que dicha selección sea realizada de forma aleatoria. Finalmente comprobaremos si los dos efectos descritos anteriormente son consecuencia de las limitaciones que presentan los métodos de expansión actuales. Para desarrollar esta tarea se aplicará un nuevo método de expansión que maximiza el ratio de mejora que aparece al expandir ya que se obligará a que en todos los casos la calidad de la consulta expandida mejore a la de la consulta original.

5.4.1. Expansión selectiva aleatoria

Los resultados observados reflejan que el uso de precisión media como medida de calidad para decidir en qué caso se debe expandir no produce el rendimiento deseado. Este hecho deja entrever la no idoneidad de la aplicación de dicha medida a tareas de expansión selectiva de consultas. Para comprobar de forma experimental si en realidad la precisión media aporta la información necesaria para estimar cuando una consulta debe ser expandida o no, se desarrolla el siguiente experimento: se compara el rendimiento de la expansión selectiva en base a la precisión media, con un proceso de expansión selectiva aleatorio. Así, la selección de aquellas consultas que serán

expandidas se realizará de forma aleatoria sobre el total de consultas. El número de consultas a expandir según cada tipo de consultas es equivalente al utilizado en los experimentos anteriores. Este número, por cada tipo y colección, aparece reflejado en la Tabla 5.5.

Tabla 5.5: Tamaños de los grupos de consultas según su grado de dificultad que resultan de la aplicación del algoritmo de las k-medias en base a la precisión media.

	Fáciles	Medias	Difíciles	Total
Robust2004	31	97	121	249
WT10g	8	31	58	97
GOV2	42	49	58	149

Así por ejemplo, para evaluar el método de expansión selectiva aleatoria en el grupo de consultas consideradas *'Fáciles´* dentro de las consultas de *Robust2004*, se procede seleccionando aleatoriamente 31 consultas sobre el total de 249 de forma equiprobable. Esta selección se realiza mediante un muestreo sin reemplazamiento y se repite un total de 10000 veces, considerándose la media de los 10000 resultados como el resultado final. Debe destacarse que en todos los casos la desviación de los 10000 resultados no supera el valor de 0.004.

Los resultados de la expansión selectiva aleatoria aparecen en la Tabla 5.6. Para facilitar la comparación de los resultados, en la misma tabla se incluyen entre paréntesis los valores obtenidos al realizar la expansión selectiva usando como criterio de expansión la precisión media. Estos últimos valores corresponden a los mostrados en la Tabla 5.3.

El grado de coincidencia que aparece con la aplicación de ambos métodos es tan elevado que lleva a concluir que expandir selectivamente usando la precisión media como estimador es muy similar a expandir selectivamente de forma aleatoria, no apreciándose diferencias sustanciales. Por tanto, se puede inferir que con los métodos actuales de expansión, el valor de precisión media que obtiene una consulta es prácticamente irrelevante en relación al valor que obtendrá al ser expandida. Este hecho es completamente contrario al enfoque actual en el campo de los métodos de predicción de la consulta en el que el objetivo principal se centra en predecir el valor de precisión media de una consulta. Sin embargo no parece clara la relación que habitualmente se presume entre la precisión media de una consulta y su capacidad de mejora al ser expandida.

5.4.2. Expansión mediante programación evolutiva

A partir de los resultados obtenidos en la sección anterior, se concluía que la calidad de una consulta en términos de precisión media no indica

Tabla 5.6: *MAP* que se obtiene al realizar la expansión selectiva de forma aleatoria. Entre paréntesis aparece el valor de *MAP* que se obtiene al usar la precisión media de cada consulta para decidir en qué caso se debe expandir.

	Fáciles	Medias	Difíciles
Robust2004	0.2468(0.2432)	0.2588(0.2633)	0.2631(0.2622)
WT10g	0.2085(0.2017)	0.2075(0.2104)	0.2063(0.2100)
GOV2	0.2958(0.2945)	0.2946(0.2971)	0.2937(0.2926)

si dicha consulta mejorará al ser expandida con un método clásico de expansión. Sería posible argumentar que este hecho no es tanto consecuencia del uso incorrecto de la precisión media como estimador, sino de las propias características del método de expansión que se aplique.

Así y con el objetivo de analizar todos los aspectos del problema aquí planteado, en esta sección se desarrollará un método de expansión con características óptimas. Consideraremos óptimo a este método de expansión ya que es capaz de incrementar la calidad de la respuesta a una consulta siempre que ésta sea expandida. El método aquí planteado difiere en gran medida de lo que podemos encontrar en la literatura ya que con los métodos actuales no es posible asegurar que la expansión beneficiará a todas y cada una de las consultas propuestas. Con el uso de este método de expansión en un entorno de expansión selectiva, se comprobará si la información suministrada por el valor de precisión media de una consulta es o no significativo a la hora de indicar si dicha consulta debe ser expandida. Como consecuencia se comprueba experimentalmente si la limitación observada con la aplicación de la medida de precisión media en relación a la expansión selectiva de consultas, aparece de forma independiente al método de expansión aplicado.

Podría esperarse que aquellas consultas con una precisión media superior mejoren su rendimiento al ser expandidas en mayor medida que aquellas con menor valor de precisión media, y por lo tanto no observar el comportamiento que aparece con la expansión selectiva aleatoria usando la divergencia *KL* como método de expansión.

Preliminares

El método de expansión desarrollado en este apartado tiene ciertas similitudes con un trabajo anterior realizado por Robertson et. al (Cao et al. (2008)). En este trabajo Robertson plantea la posibilidad de calcular un umbral superior ideal para la expansión de consultas, analizando el efecto de introducir términos candidatos de manera individual en la consulta original. Así, según el algoritmo planteado por Robertson se añaden de uno en uno términos candidatos en una consulta. En caso de que un término incremente el valor de precisión media, se añadirá a la consulta expandida. Este proceso se repite para cada uno de los términos candidatos existentes

en cada consulta. Como resultado se obtiene un valor óptimo de precisión media en términos de expansión para cada consulta, así como el conjunto de términos candidatos óptimo para dicha consulta.

La principal diferencia entre el enfoque propuesto por Robertson y el aplicado aquí es que la selección de términos no se basa en su efecto individual en la expansión, sino que se busca aquella combinación de términos candidatos que incremente en un grado mayor el valor de precisión media de la consulta expandida. Es decir, se busca la combinación de términos candidatos a partir de los documentos considerados, de forma que la combinación entre los términos originales y los candidatos maximice el valor de precisión media. Al añadir a la consulta original una combinación de términos, en vez de términos de manera individual, se tiene en cuenta el efecto que los términos candidatos tienen entre sí y su efecto sobre el valor de precisión media que obtendrá la consulta final.

Con el enfoque original de Robertson, una vez que un término candidato ha sido añadido a una consulta, éste no puede ser eliminado y por tanto no se tiene en cuenta el efecto que puede producir sobre el rendimiento de la consulta expandida la aparición de este término en combinación con otros que se incluyan posteriormente.

Para la selección de aquella combinación de términos candidatos óptima dada una consulta nos apoyaremos en el uso de algoritmos evolutivos. Este tipo de enfoque algorítmico ha sido utilizado con anterioridad en el campo de la recuperación de la información (Cordón et al. (2003)). De forma específica se han utilizado en el campo de la expansión de consultas en Robertson y Willet (1996); Yang y Korfhage (1994); Sanchez et al. (1995); Lopez-Pujalte et al. (2002); Horng y Yeh (2000), donde se usaban para estimar el peso asignado a cada término de la consulta expandida, o bien en Cordón et al. (2002); Araujo y Pérez-Agüera (2008) para seleccionar nuevos términos para la consulta original.

Con la disponibilidad de los juicios de relevancia de cada consulta, podemos diseñar un algoritmo genético por cada una de las consultas, de forma que se busque la combinación de términos óptima mediante el uso de una función de ajuste que optimice el valor de precisión media de cada consulta usando dichos juicios de relevancia. Puesto que el número de términos candidatos considerados para la expansión es elevado (40 términos), la búsqueda de la combinación óptima de términos no puede ser realizada de forma exhaustiva, ya que se debería evaluar hasta $2^{40} \approx 10^{12}$ combinaciones diferentes. Esto hace necesario la utilización de algún método de optimización aproximado como los algoritmos evolutivos.

Diseño del algoritmo evolutivo de expansión

Por cada una de las consultas y en base a los diez documentos devueltos en primer término haciendo uso de la consulta original, se seleccionan los

40 términos con mayor divergencia *KL*. Con el objetivo de simplificar el algoritmo evolutivo, no se aplica ningún tipo de técnica de repesado a los términos candidatos, por lo que los términos originales de la consulta y los términos de expansión tienen el mismo peso. Sin el uso de pesos para cada uno de los términos, se establece una representación binaria del problema de forma natural. Los individuos que representan las soluciones candidatas al problema son cadenas de ceros y unos, donde uno representará que dicho término será incluido en la consulta expandida y cero la omisión de dicho término. Aquellos términos que aparecieran en la consulta original siempre serán incluidos en la consulta expandida.

En la Tabla 5.7 aparecen los parámetros del algoritmo evolutivo de selección óptima de términos de expansión.

Tabla 5.7: Parámetros del algoritmo evolutivo de selección óptima de términos de expansión.

Parámetro	Valor
Longitud de cruce	40
Población	200
Ratio de cruce	0.5
Ratio de mutación	0.05
Número máx. evaluaciones	200

La población inicial se genera de forma aleatoria, con un tamaño de 200 individuos. El algoritmo genético se ejecutará un máximo de 200 generaciones, aplicando el operador de cruce monopunto, y mutaciones aleatorias de un bit. La probabilidad de cruce se fija en 0.5 y la de mutación en 0.05.

Una vez recuperado el conjunto de documentos, se realiza la extracción de términos utilizando el método basado en *KLD* descrito anteriormente. Posteriormente este conjunto de términos candidatos junto a la consulta original sirven de entrada al algoritmo genético. La función principal del algoritmo genético consiste en buscar aquella combinación entre la consulta original y un subconjunto de los términos candidatos tal que la precisión media se maximice. Para realizar esta tarea se implementa una función de ajuste que haciendo uso de los juicios de relevancia generados, permite obtener el valor de precisión media por cada combinación de términos generada por el algoritmo genético.

Para los resultados que se muestran a continuación solo se tuvieron en cuenta las 150 primeras consultas de *Robust2004*, esto es de la 301 a la 450.

En la Tabla 5.8 aparecen los resultados que se obtienen al expandir de manera sistemática el total de las consultas, con los términos proporcionados por el algoritmo genético, respecto a no realizar la expansión. Como se puede observar el rendimiento mostrado por el método de expansión es considerablemente superior al basado en *KLD*, ya que se alcanza un incre-

mento en el valor de precisión media que dobla al mostrado con el uso de las consultas originales.

Tabla 5.8: Resultados comparativos entre consulta base y consulta expandida, así como los resultados de *SQE* óptimo y peor caso con *SQE* para las consultas de la 301 a la 450.

	Robust2004
Original	0.2131
Expandida	0.4518
SQE_{Max}	0.4518
SQE_{Min}	0.2131

La precisión media que se obtiene con la aplicación de la expansión selectiva por grupos, aparece en la Tabla 5.9. De nuevo y como ocurría con la expansión basada en *KLD* los resultados indican la idoneidad de expandir aquel grupo que contiene mayor número de consultas. En este caso, el grupo considerado *'Difíciles'* que contiene 86 consultas obtiene una precisión media de 0.34 respecto al grupo considerado *'Fáciles'* que contiene 15 consultas, como aparece en la Tabla 5.9.

Tabla 5.9: *MAP* que se obtiene al realizar la expansión selectiva por grupos utilizando como criterio de expansión la precisión media que obtiene cada consulta. Entre paréntesis aparece el número de consultas consideradas en cada grupo.

	Fáciles(15)	Medias(49)	Difíciles(86)
Robust2004	0.2332	0.3014	0.3430

Finalmente se realiza la expansión selectiva de manera aleatoria. Los resultados aparecen en la Tabla 5.10, observándose un comportamiento equivalente al observado con el uso de *KLD*. Esto es, no aparece una mejora de rendimiento en la expansión selectiva con el uso de información real de la calidad de la consulta respecto a realizarlo de manera completamente aleatoria. Este hecho viene apoyado por el valor de correlación, $\tau = 0{,}12$, que aparece entre la precisión media de las consultas sin expandir y el incremento en dicho valor de las consultas tras su expansión.

Tabla 5.10: *MAP* que se obtiene al realizar la expansión selectiva de forma aleatoria. Entre paréntesis aparece el número de consultas consideradas en cada grupo.

	Fáciles(15)	Medias(49)	Difíciles(86)
Robust2004	0.2370	0.2911	3499

Ambos hechos indican que independientemente del método de expansión aplicado, este se verá afectado por la poca significancia de una medida de calidad como la precisión media a la hora de decidir en qué casos una consulta debe o no ser expandida.

La conclusión final del conjunto de resultados aquí plasmados indica la falta de idoneidad de la medida de precisión media como estimador en un marco de expansión selectiva. Este hecho es de gran importancia ya que la mayoría de los métodos de predicción se centran en estimar la precisión media, sin analizar si esta medida de calidad tiene sentido dentro del marco de actuación donde se desean utilizar las predicciones. Para el caso de la expansión selectiva una medida que podría ser más adecuada es la precisión a 10 (*P@10*), ya que los términos candidatos para la expansión provienen de los primeros 10 documentos que son devueltos con la consulta original. En la siguiente sección se intentará dar respuesta a la pregunta que plantea dicho escenario, esto es, si *P@10* es una medida de calidad más idónea para el caso de la expansión selectiva.

5.4.3. Precisión a 10 como estimador

Los resultados anteriores pueden sugerir la no utilidad de los métodos de predicción para aumentar el rendimiento en la tarea de expansión de consultas. Sin embargo, sería deseable comprobar experimentalmente si esta falta de rendimiento es consecuencia de la aplicación de los métodos de predicción o como parece intuirse con los resultados mostrados, es debido al uso de la precisión media como estimador a la hora de realizar la expansión. Por ejemplo parece plausible que para la tarea de expansión un método de predicción debería centrarse en estimar el valor de precisión a diez (*P@10*) en lugar del de *AP*. *P@10* mide la precisión entre los diez primeros documentos recuperados por un sistema y generalmente la expansión de consultas se realiza en base a estos diez primeros documentos.

Para comprobar la validez de la medida de *P@10* como estimador en un entorno de expansión selectiva, se plantea un experimento similar a los realizados previamente. Así, se generan grupos de consultas en base a su calidad, como se ha hecho de forma habitual, pero usando en este caso la medida *P@10*. Los detalles de los grupos resultantes se muestran en la Tabla 5.11. Como era esperable la estructura de los grupos y el número de elementos que contienen sufre ciertas modificaciones respecto a los obtenidos con la precisión media, consecuencia de que los valores que puede tomar una consulta en relación a la medida *P@10* se limitan a 11 según la siguiente ecuación:

$$P@10 = n/10$$

donde $n \in [0, 10]$ y $n \in \mathbb{N}$.

Los resultados de aplicar la expansión selectiva con los grupos previamente generados aparecen en la Tabla 5.12. En este caso, se observa una leve mejoría expandiendo aquellas consultas consideradas *'Fáciles'* respecto al resto de grupos. Este hecho se repite para el conjunto de las tres colecciones.

Aunque las diferencias que se observan no son importantes, sí indican la idoneidad de la expansión para aquellas consultas con valores superiores de *P@10*. Destaca el hecho de que con el uso de *P@10* como estimador se observa que los mejores resultados no se obtienen con la expansión del grupo que contenga más consultas, al contrario de lo que ocurría con la precisión media como estimador.

Los resultados comparativos entre la expansión selectiva, usando como estimador *P@10*, y la expansión sistemática de todo el conjunto de consultas aparecen en la Tabla 5.2. En base a los resultados mostrados, se puede observar un leve incremento o resultados similares en el rendimiento con la aplicación de la expansión selectiva para el caso de las colecciones *WT10g* y *GOV2*. Más específicamente en el caso de *WT10g* la expansión selectiva produce un valor de MAP=0.2151 respecto a 0.2046 con la expansión de todas las consultas. Para el caso de *GOV2* se pasa de 0.3078 a 0.3081. Sin embargo, en el caso de las consultas de la tarea *Robust2004*, no se aprecia el mismo nivel de rendimiento con la aplicación de la expansión selectiva. Aún así, el grado de mejora que se observa en *WT10g* y *GOV2* no aparecía al utilizar la precisión media como estimador. Este hecho sugiere la mayor idoneidad de *P@10* respecto a *AP*.

Tabla 5.11: Tamaños de los grupos de consultas según su grado de dificultad que resultan de la aplicación del algoritmo de las k-medias en base a *P@10*.

	Fáciles	Medias	Difíciles	Total
Robust2004	87	43	119	249
WT10g	35	40	22	97
GOV2	57	38	54	149

Tabla 5.12: *MAP* que se obtiene al realizar la expansión selectiva por grupos utilizando como criterio el valor real de *P@10* de cada una de las consultas.

	Fáciles	Medias	Difíciles	Todas
Robust2004	0.2652	0.2532	0.2509	0.2863
WT10g	0.2151	0.1998	0.2073	0.2046
GOV2	0.3081	0.2918	0.2843	0.3078

Finalmente se quiere comprobar experimentalmente si existen diferencias reales entre este último método de expansión guiado por *P@10*, respecto a una expansión aleatoria usando los tamaños de los grupos de consultas

indicados por la misma medida. Los resultados del experimento aleatorio aparecen en la Tabla 5.13, donde se puede comprobar como con la selección aleatoria de consultas para su expansión basada en *P@10* ocurre algo similar a lo que ocurría con el uso de *AP*. Es decir, se obtienen los mejores resultados si se expande aquel grupo que contenga mayor número de consultas, como es el caso de las *'Difíciles'* para *Robust2004* y las consideradas *'Fáciles'* para *GOV2*. En el caso de *WT10g* no se observan apenas diferencias entre la expansión por distintos grupos. Este hecho muestra claramente que existen diferencias significativas entre una expansión selectiva aleatoria y la que utiliza *P@10* como estimador, al contrario de lo observado con el uso de la precisión media como estimador.

Tabla 5.13: *MAP* que se obtiene al realizar la expansión selectiva de forma aleatoria usando los grupos de consultas obtenidos usando la medida *P@10*.

	Fáciles	Medias	Difíciles
Robust2004	0.2570	0.2490	0.2628
WT10g	0.2072	0.2071	0.2078
GOV2	0.2957	0.2932	0.2953

Finalmente se muestra de forma gráfica la relación existente entre el valor de *P@10* que obtiene una consulta y su potencial incremento en términos de precisión media cuando ésta ha sido expandida. Así en la Figura 5.1, se observa como evoluciona en términos de rendimiento la calidad de las consultas expandidas, dependiendo de la calidad de las consultas originales a expandir. En rojo se muestra la precisión media mientras que *P@10* aparece en azul. Cada uno de los gráficos corresponde con una de las colecciones de test.

El eje inferior representa el valor límite máximo que tienen las consultas en términos de *P@10* y *AP*. El eje de ordenadas muestra el valor medio de incremento en términos de precisión media del conjunto de dichas consultas al ser expandidas.

Del conjunto de gráficas se puede destacar que no aparece una tendencia clara en la mejora de las consultas al ser expandidas dependiendo del valor de precisión media que obtenían originalmente. Sin embargo, esta tendencia aparece claramente en el caso de *P@10*, donde para valores bajos de esta medida el incremento es poco predecible. Al aumentar el valor de *P@10*, en el entorno de consultas con $P@10 \leq 0{,}6$, la tendencia que aparece es clara y describe una mejora global para el conjunto de consultas.

Este mismo hecho aparece en las Tablas 5.14, 5.15, 5.16, 5.17, 5.18 y 5.19.

En estas tablas se muestra el número de casos en los que al expandir una consulta se produce un incremento positivo de AP ($+\triangle AP$). El número de casos con incrementos positivos se muestra para distintos umbrales de AP

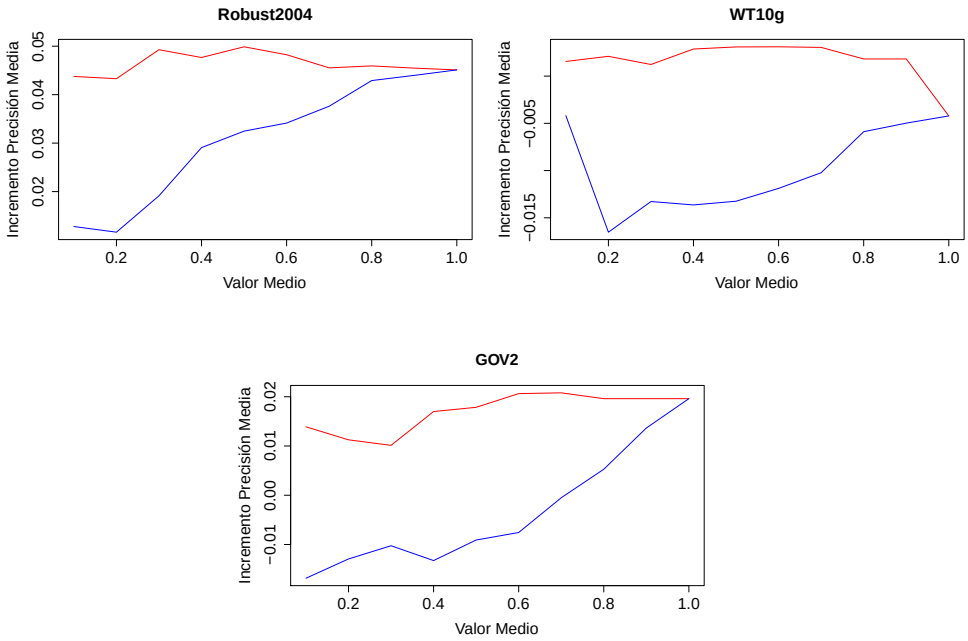

Figura 5.1: Relación entre los valores de *P@10* y *AP* de las consultas respecto a su potencial incremento de *AP* al ser expandidas para las tres colecciones de test. En rojo aparece *AP* y en azul *P@10*.

y *P@10* de las consultas sin expandir.

Las tablas aparecen en orden según las colecciones a las que se refieren. Así las Tablas 5.14 y 5.15 se refieren a *Robust2004* y muestran el porcentaje de casos positivos en relación a *AP* y *P@10* respectivamente. De la misma manera las Tablas 5.16 y 5.17 muestran los resultados para *WT10g*. Finalmente, las Tablas 5.18 y 5.19 muestran los casos positivos para *GOV2*.

En el caso de los datos obtenidos para las consultas de *Robust2004*, se observa que el número de casos positivos no solo no crece a mayores valores de *AP* sino que decrece. Por ejemplo, para el total de las consultas ($AP > 0$) el porcentaje corresponde a un 61 % mientras que para aquellas consultas que obtienen un valor de *AP* superior a 0.5 el porcentaje decae hasta un 58 %.

Sin embargo para el caso de *P@10* en esta misma colección, el número de casos positivos crece de forma monótona al ir incrementándose el valor de *P@10*.

Este comportamiento se muestra de forma mucho más clara al observar los resultados en la colección *WT10g*, donde el porcentaje de casos positivos decrece de manera casi constante para el caso de *AP*, observándose el comportamiento contrario en el caso de la *P@10*.

Tabla 5.14: *Robust2004*. Número de casos en los que el valor de precisión media se incrementa con la expansión de consultas para distintos umbrales de precisión media.

AP	> 0	$> 0{,}1$	$> 0{,}2$	$> 0{,}3$	$> 0{,}4$	$> 0{,}5$	$> 0{,}6$	$> 0{,}7$	$> 0{,}8$	$> 0{,}9$
Casos	249	171	128	81	51	29	19	11	5	1
$+\triangle AP$	152	106	82	51	35	17	10	7	2	0
%	61 %	62 %	64 %	63 %	69 %	58 %	53 %	64 %	40 %	0 %

Tabla 5.15: *Robust2004*. Número de casos en los que el valor de precisión media se incrementa con la expansión de consultas para distintos umbrales de precisión a 10.

P10	> 0	$> 0{,}1$	$> 0{,}2$	$> 0{,}3$	$> 0{,}4$	$> 0{,}5$	$> 0{,}6$	$> 0{,}7$	$> 0{,}8$	$> 0{,}9$
Casos	221	200	167	130	105	87	60	38	17	6
$+\triangle AP$	140	130	111	91	73	62	48	31	15	6
%	63 %	65 %	66 %	70 %	70 %	71 %	80 %	81 %	88 %	100 %

Tabla 5.16: *WT10g*. Número de casos en los que el valor de precisión media se incrementa con la expansión de consultas para distintos umbrales de precisión media.

AP	> 0	$> 0{,}1$	$> 0{,}2$	$> 0{,}3$	$> 0{,}4$	$> 0{,}5$	$> 0{,}6$	$> 0{,}7$	$> 0{,}8$	$> 0{,}9$
Casos	95	64	39	26	16	8	4	3	1	1
$+\triangle AP$	46	32	18	12	5	1	0	0	0	0
%	48 %	50 %	46 %	46 %	31 %	12 %	0 %	0 %	0 %	0 %

Tabla 5.17: *WT10g*. Número de casos en los que el valor de precisión media se incrementa con la expansión de consultas para distintos umbrales de precisión a 10.

P10	> 0	$> 0{,}1$	$> 0{,}2$	$> 0{,}3$	$> 0{,}4$	$> 0{,}5$	$> 0{,}6$	$> 0{,}7$	$> 0{,}8$	$> 0{,}9$
Casos	82	62	45	32	22	17	13	10	2	1
$+\triangle AP$	42	32	25	19	16	14	11	8	2	1
%	51 %	52 %	55 %	59 %	72 %	82 %	85 %	80 %	100 %	100 %

Finalmente en el caso de *GOV2* se observa la misma tendencia que en los casos anteriores pero de forma mucho más suave. Así para el caso de *AP*, el número de casos positivos no cae bruscamente, como en *WT10g*, sino que se mantiene de forma más o menos estable independientemente del valor de precisión media que obtienen las consultas consideradas. La única salvedad aparece con el número de casos positivos para consultas con *AP* altos, por ejemplo 0,7, donde el porcentaje cae hasta un 25 % pero más como consecuencia del bajo número de consultas que superan dicho umbral.

En el caso de *P@10* y para esta misma colección se observa una tendencia de crecimiento casi monótona, más suave que la observada para *WT10g* y muy similar a la que aparece en el caso de *Robust2004*.

Tabla 5.18: *GOV2*. Número de casos en los que el valor de precisión media se incrementa con la expansión de consultas para distintos umbrales de precisión media.

AP	> 0	$> 0{,}1$	$> 0{,}2$	$> 0{,}3$	$> 0{,}4$	$> 0{,}5$	$> 0{,}6$	$> 0{,}7$	$> 0{,}8$	$> 0{,}9$
Casos	147	116	91	69	45	23	10	4	0	0
$+\triangle AP$	81	67	57	47	30	14	5	1	0	0
%	55 %	58 %	59 %	68 %	66 %	60 %	50 %	25 %	NA	NA

Tabla 5.19: *GOV2*. Número de casos en los que el valor de precisión media se incrementa con la expansión de consultas para distintos umbrales de precisión a 10.

P10	> 0	$> 0{,}1$	$> 0{,}2$	$> 0{,}3$	$> 0{,}4$	$> 0{,}5$	$> 0{,}6$	$> 0{,}7$	$> 0{,}8$	$> 0{,}9$
Casos	124	111	105	92	86	79	66	54	40	22
$+\triangle AP$	76	73	70	63	62	58	51	42	31	17
%	61 %	66 %	67 %	68 %	72 %	73 %	77 %	78 %	78 %	77 %

5.5. Conclusiones

Del análisis realizado en este capítulo se puede concluir que la precisión media no es un estimador adecuado para el caso de la expansión de consultas selectiva. Por tanto el desarrollo de métodos de predicción para la expansión selectiva debe poner centrarse en medidas más adecuadas a esta tarea como por ejemplo la medida *P@10*. Una primera aproximación a la predicción de la calidad de consultas en base a *P@10* aparece en el trabajo publicado por Zhou y Croft (2006).

Los resultados aquí mostrados no dependen del método de predicción específico en uso ya que los resultados se han obtenido usando los valores reales que obtienen las consultas en términos de *AP* y no las predicciones. Este hecho sugiere que el comportamiento de cualquier otro método de pre-

dicción enfocado a predecir AP mostraría similares limitaciones a la hora de ser aplicado a un entorno de expansión selectiva de consultas.

Este hecho no hace más que demostrar que la utilidad de los métodos de predicción depende en gran medida del entorno de aplicación y que por tanto el objetivo a estimar depende del uso que se le pretenda dar al método de predicción. Una vez demostrado el potencial de los métodos de predicción a la hora de predecir AP sería de desear que el desarrollo de este tipo de métodos se desligue del valor de precisión media que obtiene una consulta y dando un paso más allá, establezca su enfoque en aquellas tareas a las que se desee aplicar.

I am turned into a sort of machine for observing facts and grinding out conclusions.

Charles Darwin.

Capítulo 6

Conclusiones

El trabajo desarrollado en esta tesis abarca tres partes fundamentales dentro del área de la predicción de la calidad de las consultas: métodos de predicción, evaluación del rendimiento de los métodos de predicción y la aplicación de las técnicas de predicción al escenario de la expansión de consultas selectiva.

Inicialmente este trabajo se centra en el desarrollo de un nuevo método de predicción. Ésta nueva aproximación se basa en el cálculo de la desviación estándar que muestran los valores de relevancia asignados por una función de ranking. La calidad de las predicciones que resultan con esta nueva propuesta son similares al rendimiento de otros métodos actuales, contando con la principal ventaja de evitar la aplicación de técnicas complejas para obtener las predicciones.

En segundo lugar se desarrolla un nuevo marco de evaluación, cuyo objetivo es superar algunas de las limitaciones que aparecen como consecuencia de una evaluación basada únicamente en coeficientes de correlación.

Finalmente, se evalúa la utilidad de los métodos de predicción en el escenario de la expansión selectiva de consultas en ausencia de información del usuario. Este escenario es utilizado asiduamente como caso de uso típico, donde la aplicación de métodos de predicción implicaría una mejora en el rendimiento, frente a la expansión sistemática de todas las consultas. Sin embargo, es difícil encontrar en la literatura relacionada integraciones exitosas de métodos de predicción en dicho escenario. Partiendo de los pobres resultados que se obtienen, se realiza un análisis de la aplicación de los métodos de predicción en este caso de uso, con el objetivo de dar explicación al bajo rendimiento encontrado.

Predicción basado en la desviación estándar

Una de las principales conclusiones que se extrae a partir del método de predicción propuesto en el Capítulo 3, es el hecho de que la calidad de una consulta puede ser estimada en base a la desviación que muestran los

pesos o valores de relevancia asignados. Así, la desviación estándar sirve como estimador de la habilidad de una función de ranking, para discriminar documentos relevantes de los no relevantes, usando para ello los valores asignados a cada documento.

Existen, dos características principales que pueden afectar el rendimiento que muestra la desviación estándar como método de predicción: el número de documentos utilizados para medir la desviación y el uso de algún proceso de normalización sobre los valores de relevancia. El proceso de normalización se suele aplicar, para facilitar la comparación de valores de relevancia que obtienen diferentes consultas. Los resultados experimentales obtenidos indican que el número de documentos k, incluidos en la medida de la desviación, afecta de forma importante a la calidad de las predicciones. Este efecto, es consecuencia de la aparición de la denominada "larga cola", formada mayoritariamente por documentos no relevantes. Por el contrario, la aplicación de procesos de normalización no ha mostrado un importante efecto en la calidad de las predicciones que se obtienen.

El método de predicción propuesto, muestra un rendimiento similar en comparación a otros métodos clásicos que aparecen en la literatura. El aspecto principal en el que destaca esta nueva técnica es su simplicidad y el bajo coste computacional que requiere. Ambas características, permitirían la aplicación de esta aproximación a entornos de búsqueda sin que conllevara un coste añadido muy elevado, aparte del requerido para la recuperación de documentos, que es común para todos los métodos *Post-Retrieval*.

Finalmente, debe destacarse que el enfoque propuesto en esta tesis ha tenido cierto impacto en el campo de la predicción de calidad de consultas. Un enfoque prácticamente equivalente al introducido en esta tesis (Perez-Iglesias (2009); Pérez-Iglesias and Araujo (2009)), fue presentado casi de forma simultánea por Shtok et al. (2009).

Shtok propone utilizar la desviación estándar de los valores de relevancia, pero en este caso normalizados por un estadístico dependiente de la colección donde se vaya a evaluar el rendimiento del método de predicción. Como es de esperar, los resultados que se obtienen con esta última aproximación son muy similares a los resultantes con la propuesta descrita en el Capítulo 3.

Recientemente, Cummins et al. (2011a) han publicado un nuevo método de predicción, que en realidad puede verse como una extensión del enfoque presentado en esta tesis. La principal diferencia, entre ambas aproximaciones, radica en el número de documentos que se utilizan a la hora de calcular la desviación estándar. Este número viene definido en función del valor que obtiene el documento devuelto en primer lugar, es decir el valor de relevancia más elevado. Así, Cummins propone calcular la desviación estándar utilizando solo aquellos documentos, cuyo valor de relevancia supere la mitad del valor que haya obtenido el documento recuperado en primer lugar. Los resultados que se obtienen con este método superan ligeramente a los presentados en esta tesis.

En otra publicación realizada por el mismo autor (Cummins et al. (2011b)), y en base a los resultados que obtiene, el autor destaca el rendimiento que muestran los métodos de predicción que hacen uso de la desviación estándar: *"The best predictors tend to be the ones based on standard deviations..."*.

Una conclusión similar a la que llega Claudia Hauff al analizar este tipo de aproximaciones en su tesis doctoral (Hauff (2010), pág. 66-67). Así, la autora destaca el rendimiento de esta aproximación *...retrieval score based methods achieve similar or higher correlations than the evaluated document content based approaches (including Clarity Score).*, con la principal ventaja de su simplicidad *"...the very low complexity, compared to approaches relying on document content, or document and query perturbations."*. Así mismo, la autora destaca como principal limitación de la predicción basada en la desviación estándar, la dependencia de esta técnica en la disponibilidad de los valores de relevancia: *"the reliance on retrieval scores can also be considered a drawback, as such approaches require collaborating search systems that make the retrieval status values available."*.

Propuesta de marco de evaluación

La evaluación de los métodos de predicción de la calidad de una consulta no había sido analizada en profundidad, sin embargo ya había sido destacado anteriormente (Hauff (2010), pág. 149) la importancia de explorar nuevas direcciones en la evaluación de los métodos de predicción. Como parte del desarrollo de esta tesis se ha realizado un profundo análisis del marco de evaluación actual con el objetivo de proponer nuevos métodos de evaluación más informativos y que superaran algunas de las limitaciones que existen en la actualidad. De esta forma, en esta tesis se han destacado los principales problemas de la aplicación de los coeficientes de correlación clásicos al entorno de la predicción de calidad de consultas, así como de algunas alternativas tales como *Weighted Kendall* o sus posibles variaciones. Aparte de los problemas inherentes que surgen al aplicar una evaluación basada en la correlación, como los descritos en el Capítulo 4, el uso de coeficientes de correlación, como marco de evaluación, implica cierto grado de desacoplamiento entre el rendimiento que muestra un método de predicción y su posible aplicación en escenarios específicos. Este tipo de evaluación, proporciona una visión general del rendimiento de un método de predicción, pero no es capaz de comparar la calidad de distintos métodos de predicción cuando el interés se centra en la capacidad de predicción de estos, respecto a tipos de consultas distintas.

En base a las limitaciones descritas, se propone un nuevo marco de evaluación con el objetivo de evaluar el rendimiento que muestran los métodos de predicción, para distintos tipos de consultas según su calidad. De esta forma, se hacen explícitas las principales diferencias entre aquellos métodos que predicen con mayor acierto cuando una consulta obtendrá una respues-

ta de calidad, respecto de los métodos que muestran un mayor acierto al detectar consultas con un pobre rendimiento. Como consecuencia se facilita en gran medida la selección del método más adecuado para el marco de aplicación deseado. El marco de evaluación propuesto se basa principalmente en asumir que cada una de las consultas pertenece a un tipo único en base a una medida de calidad, como por ejemplo *AP* o *P@10*. De esta forma, se transforma el problema de evaluación de los métodos de predicción en un problema de clasificación, lo que hace posible el uso de un gran número de medidas de evaluación aplicadas habitualmente en el campo de la clasificación, como precisión, cobertura o la clásica medida-F. Este tipo de medidas pueden ser aplicados de manera parcial, es decir para grupos de consultas específicas, o bien de manera global para el total de consultas existentes. Como extensión a las medidas de evaluación clásicas en el campo de la clasificación se propone una nueva medida (*DBEM*) específica para el caso de la predicción. Esta medida no se centra en el ratio de acierto en la clasificación, como ocurre de manera habitual, sino en la penalización en que estos métodos incurren al predecir el tipo de una consulta de manera errónea.

El marco de evaluación propuesto ha sido probado, con dos aproximaciones distintas, sobre un conjunto de métodos de predicción representativos: en base a subconjuntos de consultas de interés y para el total de consultas disponibles. La evaluación de los métodos de predicción según su habilidad para clasificar consultas por grupos de dificultad, muestra ciertos detalles del rendimiento de estos que no se hacen explícitos con el uso de coeficientes de correlación. Así, se observa el bajo rendimiento de aquellos métodos basados en el estadístico *IDF*, o la tendencia general que muestran la mayoría de los métodos de predicción, a estimar con mayor precisión aquellas consultas de baja calidad.

En relación a la evaluación para el conjunto completo de consultas se observa un importante grado de correlación, en el entorno de $r = 0{,}7$, con los métodos clásicos de evaluación (*Pearson* y *Kendall*). Este hecho indica que el marco de evaluación propuesto, cuando se aplica a todo el conjunto de consultas, proporciona resultados similares a los que se obtienen con los coeficientes de correlación.

Expansión selectiva de consultas

Finalmente y en base a los resultados obtenidos en los capítulos previos, se realiza un análisis de la aplicación de métodos de predicción dentro del campo de la la expansión selectiva de consultas. En la literatura relacionada, este campo de aplicación se considera como uno de los más indicados para la aplicación de distintos métodos de predicción. A partir del análisis realizado, se comprueba experimentalmente que existe una relación limitada entre la calidad de una consulta, medida en términos de *AP*, y su potencial de mejora

en caso de que sea expandida.

Esta conclusión se alcanza de forma experimental al comprobar que el rendimiento mostrado por la expansión selectiva dirigida por el valor de precisión media, es muy similar al que se obtiene si dicha selección es guiada mediante un método totalmente aleatorio. La misma conclusión se obtiene con la utilización de un método de expansión desarrollado específicamente para esta tarea. La principal característica de este método de expansión es que asegura que en todos los casos la respuesta que obtenga la consulta expandida, superará a la obtenida por la consulta original. Este hecho sugiere que en general, la falta de rendimiento presente al realizar expansión selectiva basada en la precisión media, aparece de forma independiente del método de expansión aplicado.

Los resultados obtenidos tienen cierta importancia, ya que contradicen la suposición más extendida sobre la aplicación de métodos de predicción en el campo de la expansión de consultas. Especialmente, se observa una contradicción con las estimaciones realizadas por Claudia Hauff (Hauff et al. (2010)), en relación a qué valor de correlación es necesario para observar mejoras significativas de la expansión selectiva respecto a la expansión sistemática. Esta contradicción aparece como consecuencia de suponer como ciertos dos comportamientos relativos a la expansión, que en realidad no se corresponden con lo observado en la experimentación aquí presentada. Por un lado supone que en general aquellas consultas que obtienen un valor de AP que supera un umbral fijado manualmente, mejorarán la calidad de su respuesta al ser expandidas. Además se asume como cierto el caso contrario, es decir, que las consultas bajo otro umbral disminuyen su precisión media siempre que son expandidas.

Finalmente, se puede concluir que el uso de métodos de predicción no es la causa que produce un descenso en el rendimiento de la expansión selectiva, sino que el problema surge con la estimación de medidas de calidad, sobre los documentos recuperados, que no observan una relación directa con el caso de uso de aplicación. Así, con el uso de la medida *P@10* como estimador para decidir en qué casos se debe expandir una consulta, los resultados alcanzan un rendimiento que los aleja de la selección aleatoria, algo que no ocurría con el uso de la medida AP. Esto hecho deja intuir que la clave para aplicar adecuadamente los métodos de predicción a escenarios diversos, es no solo un método de predicción robusto, sino también la predicción de la medida más adecuada para cada uno de los escenarios propuestos.

6.1. Trabajo Futuro

Los métodos de predicción que han ido siendo propuestos a lo largo de los años, han demostrado su capacidad para obtener grados de correlación significativos con la medida de precisión media, algo que demuestra su hipotético

potencial. Sin embargo, esta capacidad de predicción no se ha conseguido trasladar a la mejora de los sistemas de recuperación de información. Es decir, no se ha observado una mejora clara en el rendimiento de estos sistemas con la aplicación de técnicas de predicción en tareas clásicas como la expansión selectiva de consultas o la meta-búsqueda. Este hecho puede ser consecuencia del estricto enfoque que se ha mantenido en cuanto a intentar predecir la precisión media, sin realizar un estudio crítico de la utilidad de dicha medida para el conjunto de tareas propuestas. Por tanto, dado que los métodos de predicción pueden ser considerados como suficientemente robustos en la actualidad, la investigación en este tipo de técnicas debe centrarse en los posibles escenarios donde su aplicación pueda suponer algún tipo de mejora.

Como ejemplo de esta nueva aproximación, se pueden citar los trabajos publicados Bellogín (2011); Bellogín et al. (2011), donde se adapta un modelo clásico de predicción como *Clarity Score* al campo de los sistemas de recomendación. En ambos trabajos se comprueba que la utilización de técnicas de predicción es capaz de mejorar la precisión de las recomendaciones suministradas por un sistema que use estas técnicas, respecto a otro que no las incluya.

La existencia de diversos escenarios de aplicación muy distintos, implica que deban ser consideradas medidas de evaluación que difieren de las contempladas en la actualidad por los métodos de predicción. Así, una línea futura de investigación debería de analizar la capacidad de predicción de los métodos actuales a la hora de estimar medidas tan distintas como *P@1*, *P@10*, precisión o cobertura. Por ejemplo, analizar si la desviación de los valores de relevancia es significativa para predecir *P@1*, o si el modelo de lenguaje de la consulta que debe ser estimado para el cálculo de *Clarity Score* dado un solo documento, como ocurriría al intentar predecir *P@1*, es suficientemente representativo.

Una aplicación de gran interés y que no ha sido tratada en profundidad, es la posibilidad de realizar evaluaciones automáticas en competiciones tipo *TREC* en ausencia de juicios de relevancia y por tanto basándonos meramente en las predicciones computadas para cada una de las consultas. Incluso sería posible combinar los métodos de predicción con otros métodos de evaluación en ausencia de juicios de relevancia propuestos previamente, como el que fue propuesto por Soboroff et al. (2001). Aquí, se plantea obtener el conjunto de documentos relevantes mediante un muestreo aleatorio del subconjunto de documentos indicado como relevantes por la mayoría de los sistemas que participan en la tarea.

Otra posible extensión a esta tesis sería la combinación de distintos métodos de predicción para mejorar la calidad de las estimaciones. Esta línea de investigación fue abierta con anterioridad por Hauff et al. (2009). Sin embargo, el uso del método de evaluación propuesto en esta tesis permite realizar combinaciones de métodos de predicción de manera más adecuada, ya que

ahora sí es posible evaluar las deficiencias o habilidades que muestra cada uno de los métodos de predicción respecto a consultas de diversas índole en cuanto a su calidad. Por tanto a priori, parece posible realizar una combinación de métodos de predicción que hayan mostrado rendimientos distintos de forma que la combinación de ellos supere las deficiencias parciales que estos métodos posean.

Una tarea que queda por resolver, ya que aún no ha sido analizada en profundidad, es la capacidad que muestran los métodos de predicción al intentar predecir consultas con distinta tipología pero esta vez no basada en la calidad de la respuesta que obtienen, sino en otro tipo de características, como la intencionalidad de la consulta o el grado de ambigüedad de éstas. La experimentación generalmente se reduce a aquellas consultas extraídas de competiciones realizadas en *TREC* y éstas son en general de tipo informativo, obviándose tipos como navegacional o transaccional. Dentro del mismo ámbito sería de interés analizar qué tipología de consultas son más fáciles de predecir en relación a su grado de ambigüedad, o dependiendo de la cobertura del asunto que describen en la colección.

La predicción de la calidad de consultas es un campo de investigación que sigue abierto, con potenciales aplicaciones en muchas tareas dentro de la recuperación de información. Muchos de los distintos métodos de predicción propuestos en los últimos tiempos, muestran una capacidad de predicción que les permitiría, a priori, ser de utilidad en muchas de estas tareas. Sin embargo, la aplicación de este tipo de técnicas no está tan extendida, ya que la comunidad se ha centrado en mejorar el rendimiento de estos sin abordar la tarea de su aplicación a escenarios concretos. Es por tanto éste uno de los aspectos donde deben centrarse los esfuerzos, de manera que se demuestre experimentalmente la utilidad de estas técnicas para mejorar la experiencia del usuario en cuanto a procesos de búsqueda.

Capítulo 7

Conclusions

This thesis is devoted to the query performance prediction subject. More specifically this work deals with three tasks within the area: prediction methods, prediction evaluation and selective query expansion based on prediction methods estimations.

The first contribution of this work is a new prediction technique proposal. This new approach is focused on computing the topic quality estimations by means of a simple calculation as the standard deviation. This approach provides a similar performance as current methods, while these last employ more complex techniques to compute their estimations.

The second proposal introduced in this thesis is related to the evaluation of query performance prediction methods. In this case, the main purpose is to develop a new evaluation framework able to provide more significant information about the performance of predictions methods, than the supplied by correlation coefficients.

Finally, the quality estimations obtained with prediction methods are applied to the selective query expansion task. The purpose of this study is to measure the potential usefulness of prediction methods within this scenario. Selective query expansion is frequently mentioned as one of the most promising potential applications of query performance prediction methods. However, in the related literature not many examples of a successful integration of these methods have been claimed. Thus, an analysis of the lack of performance in this scenario is carried out in order to achieve a better understanding of this task.

Prediction based on the standard deviation

The main conclusion obtained from the proposed prediction technique described in Chapter 3 is the suitability of standard deviation to predict the performance of a query. These predictions are obtained by simply measuring the standard deviation of the scores assigned by the ranking function to the

returned documents, after a query is posed to a search system.

The ability of current ranking functions to distinguish relevant and not relevant documents by means of the assigned scores, supports the use of standard deviation as an estimator of a query performance, since good performing queries will show a higher dispersion among its scores.

Two main aspects can affect the performance of the standard deviation as a prediction method: the number of documents included to measure the standard deviation and the use of a normalization method to standardize the scores assigned by the ranking function. The last is frequently applied in order to facilitate a fair comparison among the scores obtained by different queries.

It can be observed, through the experiments reported in Chapter 3, that the number of documents k included to measure the standard deviation, has a strong effect in relation with the quality of the estimations supplied, consequence of the "long tail" composed of the not relevant documents set. Opposite to this, the application of a normalization process do not cause a significant effect on the computed estimations.

The proposed prediction technique shows a similar performance to other approaches which appear in the related literature such as Clarity Score. The most significant advantage of the proposed method is the capacity of computing estimations without using complex techniques involving a high computational cost. This characteristic of the proposed method makes it suitable for its application to a real search system scenario.

It should be remarked that the query performance prediction technique by means of the standard deviation has caused some impact in the QPP community. Thus, a similar method to the one introduced here was proposed almost simultaneously to our approach (Perez-Iglesias (2009); Pérez-Iglesias and Araujo (2009)) by Shtok et al. (2009). Shtok, also proposed the use of the standard deviation as a prediction method, but in this case applying a scores normalization process previously. The normalization factor is based on a statistic computed from the corpus where the prediction method is tested. The results obtained with this approach are similar to the obtained with the proposed method in Chapter 3.

More recently Ronan Cummins (Cummins et al. (2011a)) has proposed an extension to the prediction method introduced in this thesis.

The main difference between both proposals is the method to select the number k of documents included to measure the standard deviation. This k size is computed based on the score received by the first retrieved document. He computes the standard deviation including only those documents with a score higher than half of the score received by the first document in the ranking list. Cummins claims slightly better results than the ones published in this thesis. In a different paper from the same author (Cummins et al. (2011b)) and according to their experiments, the authors claim *"The best predictors tend to be the ones based on standard deviations..."*.

Something similar is found by Claudia Hauff, as in her thesis (Hauff (2010), pages 66-67), she remarks the potential of the standard deviation based approach *...retrieval score based methods achieve similar or higher correlations than the evaluated document content based approaches (including Clarity Score).*, with the main advantage of *"the very low complexity, compared to approaches relying on document content, or document and query perturbations."*. As the main drawback of this approach, she remarks *"the reliance on retrieval scores can also be considered a drawback, as such approaches require collaborating search systems that make the retrieval status values available."*.

Proposed evaluation framework

In this thesis a new proposal for the evaluation of query performance prediction methods have been introduced. Although, previous works to this thesis had remarked the necessity of a more suitable approach to evaluate the quality of predictions (Hauff (2010), page 149), no research dealing with this issue has been carried out.

An important part of the work carried out in this thesis is related to the evaluation of query performance prediction methods. Therefore, an analysis dealing with the consequences of an evaluation based uniquely on correlation coefficients, or any other derived measure as Weighted Kendall is done. Based on this study, a new evaluation framework is developed. The main purpose of this new proposal is to avoid some of the drawbacks showed by the current evaluation framework, and thus providing a more informative measure of the QPP methods.

In addition to the well-known issues, which are consequence of the correlation coefficients application as an evaluation measure, described in Chapter 4, it is important to remark that a correlation value only describes the general performance of a prediction method. Thus, it is not possible to measure the accuracy of QPP methods predicting different types of queries according to their performance. In many cases, the possible application of a prediction method to a specific scenario it is not so strongly related to the global performance, but to the performance that a method shows for a specific type of queries.

Based on the drawbacks found, we introduce a new evaluation framework specifically focused on measuring the performance of prediction methods for different type of queries. Thus, this evaluation method is able to show when a prediction method is suitable for the detection of queries with a high performance against predicting queries with a low value of quality. The application of this evaluation framework allows to select the more suitable

prediction method for a specific scenario.

The proposed evaluation framework assumes that every query belongs to a class of queries. The class of a query depends on the performance showed by it using measures such as AP or P@10. Since each query is uniquely assigned to a group based on its quality, the evaluation of prediction methods can now be observed as a classification problem and thus applying any measure from this field as accuracy, recall, o F-measure. These measures can be applied to the whole set of queries or to a subset of them, focusing on the performance of the prediction method for a specific type of queries.

As an extension to the classic classification measures, the Distance Based Error Measure (DBEM) measure has been developed. This measure, opposite to others, is focused on the misclassified topics, assigning a different degree of penalty to a wrongly classified topic. The error degree is based on the distance between the real query type and the estimated one. Therefore, with this measure it is possible to observe what type of misclasifications occur more frequently with the evaluated prediction method.

The described evaluation framework is tested with a subset of the current available prediction methods. The performance of these methods is evaluated in a double fashion: by class and for the whole set of topics. Concerning the first evaluation case it is observed experimentally some details about their performance not showed by the correlation coefficients, as the low performance obtained by some of the IDF based prediction methods or the general bias showed by prediction methods to estimate with higher accuracy low quality topics.

In relation to the evaluation for the whole set of topics, a strong correlation ($r = 0{,}7$) between the results obtained with Pearson or Kendall and the global performance obtained with the proposed framework is found. This fact, implies that the proposed framework, when applied to the whole set of topics, shows similar results to the obtained with the classical correlation coefficients.

Selective query expansion

Finally, and based on the results obtained with the previous proposals, the last part of the thesis is devoted to analyze the application of query performance prediction methods to the selective query expansion scenario, when no relevance feedback from the user is available.

This use case is one of the most frequently cited to motivate the potential application of prediction methods to different scenarios. We have concluded experimentally that a low improvement can be expected from the inclusion of a prediction method in a selective query expansion scenario.

This lack of performance is due to the limited relationship between the average precision obtained by a query and the performance of the same

query after a query expansion process is applied to it.

This conclusion appears after not observing any improvement when the AP value obtained by a query is applied to decide when it should be expanded to improve its performance. This typical approach of selective query expansion is compared with a random fashion selection of the queries for expansion, obtaining both approaches almost equivalent results.

The same conclusion is observed when a different query expansion method developed specifically for this task is applied. The main characteristic of this query expansion method is the capacity to improve, in all cases, the value obtained after the expansion compared with the average precision obtained originally. This fact suggests that the lack of performance showed in a selective query expansion scenario, using average precision as estimator, is not related to the query expansion method applied.

The conclusions obtained in Chapter 5 contradict the most extended idea about the suitability of the average precision quality measure as an estimator in a selective query expansion scenario. More precisely, the obtained results are opposite to the conclusions obtained by Claudia Hauff (Hauff et al. (2010), pages 101-104), related to the correlation threshold necessary to observe an improvement with the application of prediction methods to the selective query expansion scenario.

The contradiction between both results is consequence of two assumptions taken by Hauff. In her experiments Hauff assumes that those topics with an average precision over a fixed threshold will improve their average precision after the expansion process. Simultaneously she assumes that those topics which obtain an average precision lower than another threshold will decrease their average precision when they are expanded.

Finally, the experiments are repeated but using P@10 as estimator for the selection of topics candidates for expansion. In this last case, the obtained results are far from those obtained using average precision as estimator. More important the improvement observed when P@10 is applied as an estimator to select which topics should be expanded it is far from a random selection of topics, contrary to the observed when average precision was used.

Therefore, the obtained results remark the importance of developing prediction methods focused on estimating the suitable evaluation measure, which depends on the scenario where the prediction method is applied.

7.1. Future work

The different prediction methods proposed during the last years have shown a strong capacity to obtain a significant correlation values with the average precision measure. This fact suggests the suitability of these methods

to improve the global performance of a search system.

This potential capacity has not been exploited yet to improve user search experience, since a limited improvement has been achieved with the application of query performance prediction methods to real scenarios.

This lack of improvement, despite other causes, could be a consequence of the main objective within this area, which is focused on the estimation of the average precision measure. Although, the average precision is considered as a standard quality measure in information retrieval, this measure is not as relevant as others for specific tasks, as it has been observed in relation to the selective query expansion task.

Future development on new prediction techniques should be guided by the scenario where these methods are applied, and not only by the correlation found between a prediction method and a generic measure of a response quality as average precision.

This change on the performance prediction research methodology will allow the development of new predictors or adaptations of current methods to improve the performance of specific tasks. An example of this new point of view on the development of new prediction techniques appears in the works published by Bellogín (2011) or Bellogín et al. (2011). Both works are devoted to the application of prediction techniques to improve the recommendations given to users in a generic recommender system. For this task, the authors adapt the Clarity Score method to this field, and they measure the accuracy of the prediction method in terms of how the recommendations are improved compared to a system where no predictions are applied. The obtained results are very promising and prove the utility of the application of a prediction method to this task.

Based on this new orientation on prediction techniques development, some of the current methods should be adapted in order to predict different quality measures. For instance, it should be tested if the standard deviation can be applied to estimate the quality of a response using a minimum number of documents, as it should be done for P@1 or P@10. In the same way it should be analyzed the performance showed by a method like Clarity Score, when the query language model is built from a small number of documents.

An interesting area of application of prediction methods not studied so far is the automatic generation of relevance judgments. Besides, prediction methods could be combined with different proposal for the automatic generation of relevance judgments, as it was proposed by Soboroff et al. (2001). In this work Soboroff, suggests to obtain a set of relevant documents through a random sampling from the set of documents considered relevant by the whole set of systems submitting their results to a TREC task.

Another possible extension focused on improving the accuracy of prediction methods is the combination of different prediction techniques to improve

the overall performance. This subject was previously studied by Hauff et al. (2009). However, this combination can be carried out more adequately using the information provided by the evaluation framework introduced in Chapter 4. This evaluation makes explicit the weakness and robustness of each prediction technique and thus allowing a more suitable combination.

Predictions methods are, in general, evaluated using standard TREC collections. However, the topics resolved in these collections are mainly informatives. Therefore, an analysis of the performance of predictors based on topic intentionality has not been done. For certain tasks it could be interesting to measure the accuracy of predictions on transactional or navigational topics, opposite to the current framework based on informative topics. Similarly, it could be studied the estimation quality of topics highly ambiguous or topics which are poorly represented along the document collection.

Query performance prediction keeps being an open research question with many fields of application. Some of the current techniques show a significant capacity to provide accurate estimations and these predictions can be employed in many information retrieval related tasks. The right application of these prediction techniques is the main area where the query performance prediction community should be focus on.

Apéndice A

Colecciones y sistemas

En este apéndice se recogen las distintas colecciones y detalles sobre la ejecución de las funciones de ranking empleadas a lo largo de este trabajo.

A.1. Colecciones

Todas las colecciones empleadas para la evaluación de las distintas aproximaciones presentadas en esta tesis han sido utilizadas previamente en algunas de las tareas de *TREC*. Más específicamente para el desarrollo de esta tesis se han utilizado las tres colecciones y las necesidades de información disponibles que se detallan a continuación:

A.1.1. *TREC Vol. 4 + 5*

El conjunto de documentos que contiene esta colección fueron seleccionados del *Financial Times*, el *Federal Register 94*, los *LA Times*, y el *FBIS (Foreign Broadcast Information Service)*. Este conjunto se corresponde con los denominados discos 4 y 5 de *TREC*, excluyendo los documentos del *Congressional Record*.

Esta colección se empleó para el desarrollo de la tarea *Robust2004*. En esta tarea se debía devolver un máximo 1000 documentos por cada consulta. Además, de forma excepcional, se pedía que aquellos participantes que así lo desearan enviaran el conjunto de necesidades de información ordenadas según su dificultad. El conjunto de consultas disponibles para esta tarea eran las consultas de la 301 a la 450 *(TRECs 6,7,8)* y de la 601 a la 700, siendo éstas últimas creadas explícitamente para esta tarea. Esto hacía un total de 250 necesidades de información que debían ser respondidas por los grupos participantes.

Se debe resaltar que para el conjunto de experimentos realizados en esta tesis se excluyó la necesidad informativa 672 (*"NRA membership profile"*), ya que no aparecen documentos relevantes para ella de acuerdo a los juicios

de relevancia disponibles.

A.1.2. *WT10g*

Esta colección fue construida para evaluar el rendimiento de distintas aproximaciones de recuperación en el entorno Web en 1997 y es un subconjunto de otra colección de mayor tamaño denominada *WT100G*. *WT10G* contiene 1.692,096 documentos, que ocupan 10 Gigabytes. Las necesidades de información disponibles para esta colección van de la 451 a la 500 que son las utilizadas en la tarea Web correspondiente al *TREC9*, y las que van desde la 501 a la 550, utilizadas en el *TREC2001*.

En las distintas tareas realizadas sobre la colección *WT10g*, se introdujeron consultas que debían dificultar en gran medida la recuperación de documentos a través de ellas. Así por ejemplo, aparecen ciertos errores ortográficos en las consultas 464 (*"nativityscenes"*) y 467 (*"angioplast7"*). Estos errores provocan, que sin un tratamiento específico de dichas consultas el número de documentos recuperados sea nulo. Algo similar ocurre con la consulta 531 *"who and whom"*. En este caso todos los términos de la consulta son eliminados al pasar por un filtro de palabras vacías.

Este conjunto de consultas son excluidas de los experimentos ya que se deseaba que todo el proceso experimental fuera completamente automático. Por tanto para esta colección se han tratado un total de 97 necesidades de información.

A.1.3. *GOV2*

La colección *GOV2* consiste en un conjunto de páginas Web de carácter gubernamental pertenecientes al dominio .GOV y que fueron recuperadas a principios del año 2004. Tiene un tamaño que supera los 25 millones de documentos repartidos en 426 Gigabytes. Las necesidades de información empleadas sobre esta colección corresponden a las que van desde la 701 a la 850 y que fueron empleadas en la tarea *Terabyte* desarrollada durante el año 2006. Al igual que en casos anteriores se excluye la consulta 703 (*"U.S. against International Criminal Court"*), ya que no existen documentos relevantes para dicha necesidad según los juicios de relevancia.

A.2. Sistemas de recuperación

A lo largo de la tesis se emplean un conjunto de rankings que se obtienen a partir de las necesidades de información descritas en la sección anterior. Estos rankings se obtienen a partir del campo *title* que aparece en las distintas necesidades de información. Así, este campo será utilizado como la consulta que se suministra a las distintas funciones de ranking.

Las colecciones han sido indexadas haciendo uso del marco de trabajo para recuperación de información *Lemur*[1].

Todos los documentos, antes de ser indexados, pasan por un proceso de *stemming* basado en el algoritmo de Porter (1997), además de otro proceso que consiste en la eliminación de palabras vacías. El conjunto de palabras vacías en inglés aplicado en este caso cuenta con un total de 318 palabras. Dicho conjunto puede ser consultado en la siguiente *URL*: `http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_wordsterrier`

Todas las necesidades de información han sido resueltas utilizando las funciones de ranking *BM25* y *Query Likelihood* que implementa *Lemur*.

Los parámetros utilizados para optimizar ambas funciones de ranking corresponden a los recomendados por los desarrolladores de este marco, siendo estos:

- *BM25* aplicado en *Robust2004*: $k_1 = 1{,}2$ y $b = 0{,}75$.
- *BM25* aplicado en *WT10g*: $k_1 = 0{,}2$ y $b = 0{,}6$.
- *BM25* aplicado en *GOV2*: $k_1 = 1{,}2$ y $b = 0{,}3$.
- En el caso de Query Likelihood, para las tres colecciones, se aplica un suavizado *Dirichlet* con $\mu = 1500$.

Los resultados en términos de *MAP* que se obtienen según la configuración anterior para el conjunto de colecciones aparecen en la Tabla A.1.

Tabla A.1: *MAP* que se obtiene en las distintas colecciones con los parámetros por defecto utilizados a lo largo de esta tesis.

	BM25	**QL**
Robust2004	0.2337	0.2412
WT10g	0.2028	0.2088
GOV2	0.2909	0.2882

[1]The Lemur Toolkit software, Version 4.12 of for Unix. Copyright (C) 2008 University of Massachusetts and Carnegie Mellon University.

Apéndice B

Publicaciones

En este apéndice se enumeran las distintas publicaciones que han surgido como consecuencia del trabajo desarrollado a lo largo de esta tesis. Cada publicación aparece relacionada con el capítulo al que hacen referencia. Además se incluyen otro conjunto de publicaciones, que sin formar parte de este trabajo tienen una temática afín.

B.1. Publicaciones de la tesis

- **(Capítulo 2)** Lourdes Araujo, Hugo Zaragoza, Jose R. Pérez-Agüera, and Joaquín Pérez-Iglesias. 2010. *Structure of morphologically expanded queries: A genetic algorithm approach*. Data Knowl. Eng. 69, 3 (March 2010), 279-289.

- **(Capítulo 3)** Joaquín Pérez-Iglesias. *Query performance prediction based on ranking list dispersion*. In Third BCS-IRSG Symposium on Future Directions in Information Access (FDIA 2009), FDIA 2009. eWIC-BCS, 2009.

- **(Capítulo 3)** Joaquín Pérez-Iglesias and Lourdes Araujo. 2009. *Ranking List Dispersion as a Query Performance Predictor*. In Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory (ICTIR '09) Springer-Verlag, Berlin, Heidelberg, 371-374.

- **(Capítulo 3)** Joaquín Pérez-Iglesias and Lourdes Araujo. 2010. *Standard deviation as a query hardness estimator*. In Proceedings of the 17th international conference on String processing and information retrieval (SPIRE'10). Springer-Verlag, Berlin, Heidelberg, 207-212.

- **(Capítulo 4)** Joaquín Pérez-Iglesias and Lourdes Araujo. 2010. *Evaluation of query performance prediction methods by range*. In Proceedings of the 17th international conference on String processing and

information retrieval (SPIRE'10). Springer-Verlag, Berlin, Heidelberg, 225-236.

- **(Capítulo 5)** Lourdes Araujo and Joaquín Pérez-Iglesias. 2010. *Training a classifier for the selection of good query expansion terms with a genetic algorithm.* In Proceedings of the Evolutionary Computation (CEC), 1-8

B.2. Otras publicaciones relacionadas

- José R. Pérez-Agüera, Javier Arroyo, Jane Greenberg, Joaquín Pérez-Iglesias, Víctor Fresno: *INEX+DBPEDIA: a corpus for semantic search evaluation.* WWW 2010: 1161-1162
- Joaquín Pérez-Iglesias, Víctor Fresno, José R. Pérez-Agüera: *Modelling field dependencies on structured documents with fuzzy logic.* FUZZ-IEEE 2009: 496-501
- Joaquín Pérez-Iglesias, Víctor Fresno, José R. Pérez-Agüera: *Fuzzy-Fresh: A Fuzzy Logic Approach to the Ranking of Structured Documents.* Web Intelligence 2008: 755-758

Referencias

American Consumer Satisfaction Index. ACSI. Internet portal and search engines scores, 2011.

Giambattista Amati, Claudio Carpineto, Giovanni Romano, y Fondazione Ugo Bordoni. Query difficulty, robustness and selective application of query expansion. In *European Conf. on IR Research*, pages 127–137. Springer, 2004.

Gianni Amati, Claudio Carpineto, y Giovanni Romano. Fub at trec-10 web track: A probabilistic framework for topic relevance term weighting. In *TREC*, 2001.

F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):pp. 17–21, 1973. ISSN 00031305. URL `http://www.jstor.org/stable/2682899`.

Avi Arampatzis y André van Hameran. The score-distributional threshold optimization for adaptive binary classification tasks. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 285–293, New York, NY, USA, 2001. ACM. ISBN 1-58113-331-6.

Lourdes Araujo y José R. Pérez-Agüera. Improving query expansion with stemming terms: A new genetic algorithm approach. In *EvoCOP*, pages 182–193, 2008.

Lourdes Araujo y José Pérez-Agüera. Improving query expansion with stemming terms: A new genetic algorithm approach. In Jano van Hemert y Carlos Cotta, editors, *Evolutionary Computation in Combinatorial Optimization*, volume 4972 of *Lecture Notes in Computer Science*, pages 182–193. Springer Berlin / Heidelberg, 2008.

Lourdes Araujo, Hugo Zaragoza, Jose R. Pérez-Agüera, y Joaquín Pérez-Iglesias. Structure of morphologically expanded queries: A genetic algorithm approach. *Data & Knowledge Engineering*, 69(3):279 – 289, 2010.

ISSN 0169-023X. Special Issue: 13th International Conference on Natural Language and Information Systems (NLDB 2008) - Five selected and extended papers.

Javed A. Aslam y Virgil Pavlu. Query hardness estimation using jensen-shannon divergence among multiple scoring functions. In *Proceedings of the 29th European conference on IR research*, ECIR'07, pages 198–209, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-71494-1.

Ricardo Baeza-Yates y Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1st edition, May 1999. ISBN 020139829X.

Ricardo Baeza-Yates, Vanessa Murdock, y Claudia Hauff. Efficiency trade-offs in two-tier web search systems. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 163–170, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6.

J. Beirlant, E. J. Dudewicz, L. Györfi, y E. C. Meulen. Nonparametric Entropy Estimation: An Overview. *International Journal of the Mathematical Statistics Sciences*, 6:17–39, 1997.

Alejandro Bellogín, Pablo Castells, y Iván Cantador. Predicting the performance of recommender systems: an information theoretic approach. In *Proceedings of the Third international conference on Advances in information retrieval theory*, ICTIR'11, pages 27–39, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-23317-3.

Alejandro Bellogín. Predicting performance in recommender systems. In Bamshad Mobasher, Robin D. Burke, Dietmar Jannach, y Gediminas Adomavicius, editors, *RecSys*, pages 371–374. ACM, 2011. ISBN 978-1-4503-0683-6.

Pia Borlund. The concept of relevance in ir. *J. Am. Soc. Inf. Sci. Technol.*, 54:913–925, August 2003. ISSN 1532-2882.

Chris Buckley y Ellen M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 33–40, New York, NY, USA, 2000. ACM. ISBN 1-58113-226-3.

Vannevar Bush. As we may think. *The Atlantic Monthly*, 176(1):101–108, 1945.

Guihong Cao, Jian-Yun Nie, Jianfeng Gao, y Stephen Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the*

31st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '08, pages 243–250, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-164-4.

David Carmel y Elad Yom-Tov. Estimating the query difficulty for information retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2(1):1–89, 2010.

David Carmel, Eitan Farchi, Yael Petruschka, y Aya Soffer. Automatic query wefinement using lexical affinities with maximal information gain. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, pages 283–290, New York, NY, USA, 2002. ACM. ISBN 1-58113-561-0.

David Carmel, Elad Yom-Tov, Adam Darlow, y Dan Pelleg. What makes a query difficult? In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 390–397, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7.

Claudio Carpineto, Renato de Mori, Giovanni Romano, y Brigitte Bigi. An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, 19:1–27, January 2001. ISSN 1046-8188.

Wei-Tsen Milly Chiang, Markus Hagenbuchner, y Ah Chung Tsoi. The wt10g dataset and the evolution of the web. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, WWW '05, pages 938–939, New York, NY, USA, 2005. ACM. ISBN 1-59593-051-5.

Charles L. A. Clarke, Falk Scholer, y Ian Soboroff. The trec 2005 terabyte track. In *TREC*, 2005.

Oscar Cordón, Félix de Moya Anegón, y Carmen Zarco. A new evolutionary algorithm combining simulated annealing and genetic programming for relevance feedback in fuzzy information retrieval systems. *Soft Comput.*, 6(5):308–319, 2002.

Oscar Cordón, Enrique Herrera-Viedma, Cristina López-Pujalte, María Luque, y Carmen Zarco. A review on the application of evolutionary computation to information retrieval. *Int. J. Approx. Reasoning*, 34(2-3):241–264, 2003.

Fabio Crestani, Mounia Lalmas, Cornelis J. Van Rijsbergen, y Iain Campbell. "is this document relevant?...probably": a survey of probabilistic models in information retrieval. *ACM Comput. Surv.*, 30:528–552, December 1998. ISSN 0360-0300.

Steve Cronen-Townsend, Yun Zhou, y W. Bruce. Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '02*, New York, 2002. ACM Press. ISBN 1581135610.

Steve Cronen-Townsend, Yun Zhou, y W. Bruce Croft. A framework for selective query expansion. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, CIKM '04, pages 236–237, New York, NY, USA, 2004. ACM. ISBN 1-58113-874-1.

Ronan Cummins, Joemon Jose, y Colm O'Riordan. Improved query performance prediction using standard deviation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 1089–1090, New York, NY, USA, 2011a. ACM. ISBN 978-1-4503-0757-4.

Ronan Cummins, Mounia Lalmas, Colm O'Riordan, y Joemon M. Jose. Navigating the user query space. In *Proceedings of the 18th international conference on String processing and information retrieval*, SPIRE'11, pages 380–385, Berlin, Heidelberg, 2011b. Springer-Verlag. ISBN 978-3-642-24582-4.

Susan J. Devlin, R. Gnanadesikan, y J. R. Kettenring. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62(3):pp. 531–545, 1975. ISSN 00063444.

Fernando Diaz. Performance prediction using spatial autocorrelation. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 583–590, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7.

David A. Evans y Robert G. Lefferts. Design and evaluation of the clarit-trec-2 system. In *TREC'93*, pages 137–150, 1993.

Christopher Fox. *Lexical analysis and stoplists*, pages 102–130. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1992. ISBN 0-13-463837-9.

Norbert Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255, 1992.

C. Hauff. *Predicting the Effectiveness of Queries and Retrieval Systems*. PhD thesis, Univ. of Twente, Enschede, January 2010.

Claudia Hauff, Djoerd Hiemstra, y Franciska de Jong. A survey of pre-retrieval query performance predictors. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 1419–1420, New York, NY, USA, 2008a. ACM. ISBN 978-1-59593-991-3.

Claudia Hauff, Vanessa Murdock, y Ricardo Baeza-Yates. Improved query difficulty prediction for the web. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 439–448, New York, NY, USA, 2008b. ACM. ISBN 978-1-59593-991-3.

Claudia Hauff, Leif Azzopardi, y Djoerd Hiemstra. The combination and evaluation of query performance prediction methods. In *ECIR*, pages 301–312, 2009.

Claudia Hauff, Leif Azzopardi, Djoerd Hiemstra, y Franciska de Jong. Query performance prediction: Evaluation contrasted with effectiveness. In Cathal Gurrin, Yulan He, Gabriella Kazai, Udo Kruschwitz, Suzanne Little, Thomas Roelleke, Stefan Rüger, y Keith van Rijsbergen, editors, *Advances in Information Retrieval*, volume 5993 of *Lecture Notes in Computer Science*, pages 204–216. Springer Berlin / Heidelberg, 2010. ISBN 978-3-642-12274-3.

David Hawking. Overview of the TREC-9 Web Track. In *NIST Special Publication 500-249: TREC-9*, pages 87–102, 2001.

Ben He y Iadh Ounis. Inferring Query Performance Using Pre-retrieval Predictors. In *String Processing and Information Retrieval*, pages 43–54, 2004.

Ben He y Iadh Ounis. Query performance prediction. *Information Systems*, 31:585–594, 2006.

Ben He y Iadh Ounis. Combining fields for query expansion and adaptive query expansion. *Inf. Process. Manage.*, 43:1294–1307, September 2007a. ISSN 0306-4573.

Ben He y Iadh Ounis. Studying query expansion effectiveness. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, y Chantal Soule-Dupuy, editors, *Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 611–619. Springer Berlin / Heidelberg, 2009.

Ben He y Iadh Ounis. Combining fields for query expansion and adaptive query expansion. *Inf. Process. Manage.*, 43(5):1294–1307, 2007b.

Jiyin He, Martha Larson, y Maarten de Rijke. Using coherence-based measures to predict query difficulty. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen White, editors, *Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*, pages 689–694. Springer Berlin / Heidelberg, 2008.

Jorng-Tzong Horng y Ching-Chang Yeh. Applying genetic algorithms to query optimization in document retrieval. *Inf. Process. Manage.*, 36(5): 737–759, 2000.

M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2): 81–93, June 1938.

Solomon Kullback y Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

John Lafferty y Chengxiang Zhai. Document language models, query models, y risk minimization for information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 111–119, New York, NY, USA, 2001. ACM. ISBN 1-58113-331-6.

Victor Lavrenko y W. Bruce Croft. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 120–127, New York, NY, USA, 2001. ACM. ISBN 1-58113-331-6.

Christina Lioma, Craig Macdonald, Vassilis Plachouras, Jie Peng, Ben He, and Iadh Ounis. University of glasgow at trec 2006: Experiments in terabyte and enterprise tracks with terrier. In *TREC*, 2006.

Cristina Lopez-Pujalte, Vicente P. Guerrero Bote, y Félix de Moya Anegón. A test of genetic algorithms in relevance feedback. *Inf. Process. Manage.*, 38(6):793–805, 2002. ISSN 0306-4573.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

R. Manmatha, T. Rath, y F. Feng. Modeling score distributions for combining the outputs of search engines. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–275, New York, NY, USA, 2001. ACM. ISBN 1-58113-331-6.

Christopher D. Manning, Prabhakar Raghavan, y Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715.

Massimo Melucci. On rank correlation in information retrieval evaluation. *SIGIR Forum*, 41(1):18–33, 2007. ISSN 0163-5840. doi: http://doi.acm.org/10.1145/1273221.1273223.

Massimo Melucci. Weighted rank correlation in information retrieval evaluation. In *AIRS '09: Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology*, pages 75–86, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-04768-8.

Stefano Mizzaro. Relevance: the whole history. *J. Am. Soc. Inf. Sci.*, 48: 810–832, September 1997. ISSN 0002-8231.

Josiane Mothe y Ludovic Tanguy. Linguistic features to predict query difficulty. In *Predicting Query Difficulty - Methods and Applications, SIGIR 2005*, 2005.

Javier Parapar y Álvaro Barreiro. Promoting divergent terms in the estimation of relevance models. In *Proceedings of the Third international conference on Advances in information retrieval theory*, ICTIR'11, pages 77–88, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-23317-3.

Joaquin Perez-Iglesias. Query performance prediction based on ranking list dispersion. In *Third BCS-IRSG Symposium on Future Directions in Information Access (FDIA 2009)*, FDIA 2009. eWIC-BCS, 2009.

Joaquín Pérez-Iglesias and Lourdes Araujo. Ranking list dispersion as a query performance predictor. In *ICTIR '09: Proceedings of the 2nd International Conference on Theory of Information Retrieval*, pages 371–374, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-04416-8.

Ari Pirkola y Kalervo Järvelin. Employing the resolution power of search keys. *J. Am. Soc. Inf. Sci. Technol.*, 52:575–583, May 2001. ISSN 1532-2882.

Jay M. Ponte y W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA, 1998. ACM. ISBN 1-58113-015-5.

M. F. Porter. Readings in information retrieval. chapter An algorithm for suffix stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997. ISBN 1-55860-454-5.

Jose Pérez-Agüera, Hugo Zaragoza, y Lourdes Araujo. Exploiting morphological query structure using genetic optimisation. In Epaminondas Kapetanios, Vijayan Sugumaran, y Myra Spiliopoulou, editors, *Natural Language and Information Systems*, volume 5039 of *Lecture Notes in Computer Science*, pages 124–135. Springer Berlin / Heidelberg, 2008.

Alexander M. Robertson y Peter Willet. An upperbound to the performance of ranked-output searching: optimal weighting of query terms using a genetic algorithm. *J. of Documentation*, 52(4):405–420, 1996.

S. E. Robertson. On Relevance Weight Estimation and Query Expansion. *Journal of Documentation*, 42(3):182–188, September 1986.

S. E. Robertson. On term selection for query expansion. *J. Doc.*, 46:359–364, January 1991. ISSN 0022-0418.

Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, 60:503–520(18), 2004.

Stephen Robertson. On Score Distributions and Relevance. *Advances in Information Retrieval*, 4425:40–51, 2007. URL `http://www.springerlink.com/index/10.1007/978-3-540-71496-5_7`.

Stephen Robertson y Steve Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In W. Bruce Croft and C. J. van Rijsbergen, editors, *SIGIR*, pages 232–241. ACM/Springer, 1994.

Stephen E. Robertson y Karen Sparck Jones. *Relevance weighting of search terms*, pages 143–160. Taylor Graham Publishing, London, UK, UK, 1988. ISBN 0-947568-21-2.

J. Rocchio. *Relevance Feedback in Information Retrieval*, pages 313–323. 1971.

Joseph L. Rodgers y Alan W. Nicewander. Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1):59–66, 1988. doi: 10.2307/2685263.

Ian Ruthven y Mounia Lalmas. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, 18:95–145, June 2003. ISSN 0269-8889.

G. Salton y M. E. Lesk. The smart automatic document retrieval systems an illustration. *Commun. ACM*, 8:391–398, June 1965. ISSN 0001-0782.

G. Salton, A. Wong, y C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975. ISSN 0001-0782.

E. Sanchez, H. Miyano, y J. Brachet. Optimization of fuzzy queries with genetic algorithms. application to a data base of patents in biomedical engineering. In *VI IFSA Congress, vol. II*, pages 293–296, 1995.

Mark Sanderson y Justin Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 162–169, New York, NY, USA, 2005. ACM. ISBN 1-59593-034-5.

C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27:379,423, 1948.

Anna Shtok, Oren Kurland, y David Carmel. Predicting query performance by query-drift estimation. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory*, ICTIR '09, pages 305–312, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-04416-8.

Ian Soboroff, Charles Nicholas, y Patrick Cahan. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 66–73, New York, NY, USA, 2001. ACM. ISBN 1-58113-331-6.

Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 1979. ISBN 0-408-70929-4.

Vishwa Vinay, Ingemar J. Cox, Natasa Milic-Frayling, y Ken Wood. On ranking the effectiveness of searches. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 398–404, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7.

Vishwa Vinay, Natasa Milic-Frayling, y Ingemar Cox. Estimating retrieval effectiveness using rank distributions. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 1425–1426, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3.

Ellen Voorhees. The philosophy of information retrieval evaluation. In Carol Peters, Martin Braschler, Julio Gonzalo, y Michael Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems*, volume 2406 of *Lecture Notes in Computer Science*, pages 143–170. Springer Berlin / Heidelberg, 2002.

Ellen M. Voorhees. Overview of the TREC 2004 Robust Retrieval Track. In *In Proceedings of the Thirteenth Text REtrieval Conference (TREC)*, 2004.

Ellen M. Voorhees. The trec 2005 robust track. *SIGIR Forum*, 40:41–48, June 2006. ISSN 0163-5840.

Jinxi Xu y W. Bruce Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 4–11, New York, NY, USA, 1996. ACM. ISBN 0-89791-792-8.

Jing-Jye Yang y Robert R. Korfhage. Query modification using genetic algorithms in vector space models. *Int. J. Expert Syst.*, 7(2):165–191, 1994. ISSN 0894-9077.

Emine Yilmaz, Javed A. Aslam, y Stephen Robertson. A new rank correlation coefficient for information retrieval. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 587–594, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-164-4.

Elad Yom-Tov, Shai Fine, David Carmel, y Adam Darlow. Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 512–519, New York, NY, USA, 2005. ACM. ISBN 1-59593-034-5.

ChengXiang Zhai. *Statistical Language Models for Information Retrieval*. Now Publishers Inc., Hanover, MA, USA, 2008. ISBN 1601981864.

Chengxiang Zhai y John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 403–410, New York, NY, USA, 2001a. ACM. ISBN 1-58113-436-3.

ChengXiang Zhai y John D. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR*, pages 334–342, 2001b.

Ying Zhao, Falk Scholer, y Yohannes Tsegay. Effective Pre-retrieval Query Performance Prediction Using Similarity and Variability Evidence. *Advances in Information Retrieval*, 4956:52–64, 2008.

Yun Zhou. *Retrieval Performance Prediction and Document Quality*. PhD thesis, University of Massachusetts, September 2007.

Yun Zhou y W. Bruce Croft. Ranking robustness: a novel framework to predict query performance. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 567–574, New York, NY, USA, 2006. ACM. ISBN 1-59593-433-2.

Yun Zhou y W. Bruce Croft. Query performance prediction in web search environments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 543–550, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7.