La interacción con el usuario en los sistemas de Recuperación de Información: realimentación por relevancia

Carlos G. Figuerola, Ángel F. Zazo, José L. Alonso Berrocal

Grupo de Recuperación de Información
Departamento de Informática y Automática
Facultad de Documentación. Universidad de Salamanca
C/ Francisco Vitoria, 6-16
37008. SALAMANCA - SPAIN
{figue–afzazo–berrocal}@usal.es

0.1 Resumen en español

En los sistemas de Recuperación de Información la interacción con el usuario permite la formulación de consultas más eficientes, que producen mejores resultados. La realimentación de consultas es una técnica a través de la cual el usuario, utilizando un interfaz adecuado, examina los documentos devueltos tras una primera consulta convencional y utiliza tales documentos para plantear al sistema ejemplos positivos y negativos de los documentos a recuperar. Se indican el proceso normalmente seguido en la realimentación de consultas y se muestran resultados experimentales que permiten estimar el grado de mejora en los resultados de la recuperación conseguida mediante estos sistemas.

Palabras clave: realimentación por relevancia, recuperación de información, modelo vectorial

0.2 Resumen en inglés

Interaction with the user in Information Retrieval Systems allows the formulation of more efficient queries that produce better results. Relevance feedback is a process where users identify relevant documents in an initial list of retrieval documents, and the system then creates a new query based on positive and negative examples of those documents. The user interface is an important element in the process. Usually followed process in relevance feedback technique and experimental results are shown. They allow estimating the degree of improvement in the results of the retrieval by means of

these systems.

Keywords: relevance feedback, information retrieval, vector mode

1 Introducción

Con el incremento del número de documentos en formato electrónico, se hace necesario contar con herramientas informáticas adecuadas para la recuperación de documentos. Las técnicas manuales han demostrado ser ineficaces, pues básicamente consisten en la elaboración manual de una descripción del contenido temático de cada uno de los documentos, siendo éste un trabajo costoso en tiempo, que además adolece de abundante inconsistencia (Hooper, 1965; Stubbs et al., 2000).

Por otra parte, las búsquedas conocidas como en texto libre (búsqueda de subcadenas a fin de cuentas) se han mostrado incapaces de resolver dos problemas básicos: la sinonimia y la polisemia. Por ello, la investigación en recuperación de información (RI) busca diseñar sistemas que acepten consultas en lenguaje natural y proporcionen documentos adecuados a tales consultas, ordenados según algún criterio del sistema, de acuerdo a las características de los documentos y a las necesidades informativas expresadas por el usuario en su consulta (Belkin y Croft, 1987). Uno de los modelos más conocido y difundido para el sistema de recuperación es el llamado modelo vectorial.

En el siguiente apartado se realiza una somera introducción al modelo vectorial, en la que se han destacado los conceptos más importantes. Se pasa seguidamente a describir el proceso de realimentación de consultas, y cómo puede aplicarse en el modelo vectorial. El cuarto apartado describe el experimento llevado a cabo y los resultados del mismo. Se finaliza con las conclusiones.

2 El modelo vectorial

El modelo vectorial fue definido por Salton (1968) hace ya bastantes años, y es ampliamente usado en operaciones de RI, así como también en operaciones de categorización automática, filtrado de información, etc. En el modelo vectorial se intenta recoger la relación de cada documento D_i, de una colección de N documentos, con el conjunto de las m características de la colección. Formalmente un documento puede considerarse como un vector (1) que expresa la relación del documento con cada una de esas características.

$$D_i \quad \rightarrow \quad \vec{d_i} = (c_{i1}, c_{i2}, \ldots, c_{im}) \tag{1}$$

Es decir, ese vector identifica en qué grado el documento D_i satisface cada una de las m características. En ese vector, c_{ik} es un valor numérico que expresa en qué grado el documento D_i posee la característica k. El concepto 'característica' suele concretarse en la ocurrencia de determinadas palabras o términos en el documento, aunque nada impide tomar en consideración otros aspectos.

Si se consideran los términos como características definitorias del documento, el proceso que debe seguir el sistema pasa primero por seleccionar aquellos términos útiles que permitan discriminar unos documentos de otros. En este punto debemos señalar que no todas las palabras contribuyen con la misma importancia en la caracterización del documento. Desde el punto de vista de la recuperación de información existen palabras casi vacías de contenido semántico, como los artículos, preposiciones o conjunciones, que son poco útiles en el proceso. Pero también son poco importantes aquellas palabras que por su frecuencia de aparición en toda la colección de documentos pierden su poder de discriminación. En RI todas ellas forman parte del conjunto de palabras vacías (stops words en inglés), que se eliminan en el proceso de indexación. Además de la eliminación de palabras vacías, en el proceso se pueden incluir aplicaciones léxicas como lematización o extracción de raíces, etiquetado de términos, detección de unidades multipalabra, etc.

Una vez seleccionado el conjunto de términos caracterizadores de la colección de documentos, es necesario obtener el valor de cada elemento del vector del documento. El caso más simple es utilizar una aproximación binaria, de forma que si en el documento D_i aparece el término k, el valor c_{ik} sería 1, y en caso contrario sería 0.

No obstante, una palabra puede aparecer más de una vez en el mismo documento, y además, unas palabras pueden considerarse con más peso, esto es, más significativas que otras, de forma que el valor numérico de cada uno de los componentes del vector obedece normalmente a cálculos más sofisticados que la simple asignación binaria. De otro lado, también es importante normalizar los vectores para no privilegiar documentos largos frente a otros documentos menos extensos.

$$\vec{d_i} = \frac{1}{\sqrt{\sum_{j=1}^{m} w_{ij}^2}}(w_{i1}, w_{i2}, \ldots, w_{im}) \tag{2}$$

Se han propuesto diversos métodos para calcular el peso de cada término en el vector documento (Salton y McGill, 1983; Salton y Buckley, 1988; Harman, 1992a), pero en general, para estimarlos se parte de dos ideas en cierto sentido contrapuestas: si un término aparece mucho en un documento, es importante para caracterizar ese documento. Pero si aparece en muchos documentos de la colección, no es beneficioso para distinguir un documento de los demás, dado su escaso poder discriminatorio, resultando poco útil

para la recuperación.

Para determinar la capacidad de representación de un término para un documento dado se computa el número de veces que aparece en dicho documento, obteniéndose la frecuencia del término en el documento, tf (*term frequency*).

Por otra parte, si la frecuencia de un término en toda la colección de documentos es extremadamente alta, se opta por eliminarlo del conjunto de términos de la colección (pertenece al conjunto de palabras vacías). Podría decirse que la capacidad de recuperación de un término es inversamente proporcional a su frecuencia en la colección de documentos. Esto es lo que se conoce como *idf* (*inverse document frequency*).

Así, para calcular el peso de cada elemento del vector que representa al documento se tiene en cuenta la frecuencia inversa del término en la colección, combinándola de alguna forma con la frecuencia del término dentro de cada documento. Normalmente se utiliza para ello el producto simple (Harman, 1992a).

$$w_{ij} = tf_i \cdot idf_j \tag{3}$$

Salton y Buckley (1988) experimentaron con más de 200 sistemas de cálculo de pesos, pero uno de los más utilizados viene dado por la ecuación 4, que expresa el peso del término j en el documento i.

$$w_{ij} = tf_{ij} \cdot \log \frac{N}{df_j} \tag{4}$$

donde df_j es el número de documentos en que aparece el término j.

El proceso realizado para los documentos también puede aplicarse a las consultas. Efectivamente, una consulta, Q, realizada en lenguaje natural está formada por términos, y, por tanto, puede verse como un documento más, seguramente bastante breve, aunque no siempre es así. Así pues, el mecanismo de obtención de pesos también se aplica a las consultas, para de esta manera poder disponer de representaciones homogéneas de consultas y documentos, que posibiliten obtener el grado de similitud entre ambas representaciones.

El vector representante de la consulta está formado por un vector de igual número de elementos que los vectores de los documentos. Cada elemento de ese vector expresa el grado en que cada uno de los términos de la colección representa las necesidades informativas de la persona que hace la consulta.

$$Q \quad \rightarrow \quad \vec{q} = \frac{1}{\sqrt{\sum_{j=1}^{m} p_j^2}} (p_1, p_2, \ldots, p_m) \tag{5}$$

La resolución de la consulta consiste en un proceso de establecer el grado de semejanza entre el vector consulta y el vector de cada uno de los documentos. Para una consulta determinada, cada documento arrojará un grado

de similitud determinado; aquéllos cuyo grado de similitud sea más elevado se ajustarán mejor a las necesidades expresadas en la consulta, desde el punto de vista del sistema de recuperación de información. No obstante, es el usuario el que debe decidir la relevancia de los documentos recuperados, siendo ésta una característica totalmente subjetiva del mismo.

El modo más simple de calcular la similitud entre una consulta y un documento, utilizando el modelo vectorial, es realizar el producto escalar de los vectores que los representan (ecuación 6). En esa ecuación se incluye la normalización de los vectores, a fin de obviar distorsiones producidas por los diferentes tamaños de los documentos. El índice de similitud más utilizado es el coseno del ángulo formado por ambos vectores. Para una consulta Q, el índice de similitud con un documento D_i es:

$$\mathrm{SIM}(Q, D_i) = \frac{\sum_{j=1}^{m} p_j d_{ij}}{\sqrt{\sum_{j=1}^{m} p_j^2 \cdot \sum_{j=1}^{m} d_{ij}^2}} \tag{6}$$

Hay otros métodos propuestos para calcular la similitud; un cuadro con los más importantes puede encontrarse en (Salton y McGill, 1983). Los resultados de la computación del índice de similitud entre la consulta y todos los documentos permite ordenar los resultados en orden decreciente. De esta manera se le ofrecen al usuario primero los documentos que el sistema de recuperación considera más similares con la consulta, y que pueden coincidir, o no, con lo esperado por el usuario. La relevancia es la medida que el usuario tiene para determinar si los resultados, y en qué grado, son adecuados a sus necesidades informativas. Es, por tanto, una media subjetiva.

Si el sistema dispone de un interfaz adecuado, el usuario revisará todos o parte de los documentos recuperados, y podrá determinar qué documentos considera pertinentes a sus necesidades informativas, y cuáles no. Es en este momento cuando se pueden aplicar mecanismos de recuperación como la realimentación de consultas. Esta es una técnica que permite la interacción con el usuario en los sistemas de recuperación de información, con el objetivo de aumentar la eficiencia del sistema. Se describe a continuación.

3 Realimentación de consultas

El modelo vectorial permite no sólo comparar una consulta con cada uno de los documentos de una colección, sino comparar y establecer un índice de similitud entre un documento y otro, sin más que aplicar la ecuación 6, utilizando los vectores de ambos documentos, uno actuando como consulta. De esta manera es posible efectuar recuperaciones mediante la utilización de documentos de ejemplo. El sistema no tiene más que encontrar aquellos

documentos que tienen un índice de similitud más alto con los que se le ofrecen como ejemplo. Nada impide entonces que el usuario utilice los resultados de la consulta, esto es, documentos, como nueva entrada para el sistema de recuperación.

Así, el proceso de recuperación sigue los siguientes pasos: el usuario formula una primera consulta, utilizando palabras que estima significativas o lenguaje natural describiendo sus necesidades informativas. El sistema calcula esa consulta, devolviendo los documentos cuyos índices de similitud con la consulta resultan ser más altos. El usuario examina los documentos devueltos y selecciona aquellos que estima más relevantes. El sistema toma esos documentos seleccionados y construye un nuevo vector con sus términos, el cual es utilizado como una nueva consulta, que es ejecutada devolviendo nuevos resultados.

Esto es lo que se conoce como expansión de consultas o realimentación de consultas a partir de las estimaciones de relevancia de los usuarios, y hoy día es muy utilizado en los buscadores de Internet, con la opción "páginas similares", "*more like this*", etc.

El problema básico aquí es la construcción del vector de la consulta realimentada. Este vector se construye a partir de los términos de la consulta inicial y de los términos de los documentos encontrados por esa consulta y señalados por el usuario como relevantes.

Adicionalmente, el usuario puede señalar no sólo los documentos relevantes recuperados por la consulta inicial, sino también los documentos que deben usarse como ejemplos negativos, es decir, aquéllos que en ningún caso desea. Obviamente, esos documentos no deseados contendrán también parte de los términos de la consulta inicial (en otro caso no habrían sido recuperados), junto con muchos otros. De alguna forma, al elaborar el vector de la nueva consulta realimentada, el sistema debe tener en cuenta tales términos, aunque en sentido negativo.

La manera de tener en cuenta los ejemplos positivos y negativos, partiendo de la base de que tanto unos como otros tendrán términos en común, es efectuar un cálculo adecuado de los pesos de los términos que compondrán el nuevo vector de la consulta realimentada. Uno de los esquemas de recálculo de pesos más utilizado es el conocido algoritmo de Rocchio (1971), dado por la ecuación 7. Existen otros algoritmos utilizables, algunos de los cuales pueden verse en otro trabajo de Harman (1992b).

$$\vec{q'} = \alpha\vec{q} + \frac{\beta}{n_r}\sum_{i\in rel}^{n_r}\vec{d_i} - \frac{\gamma}{n_{nr}}\sum_{j\in norel}^{n_{nr}}\vec{d_j} \qquad (7)$$

donde n_r es el número de documentos considerados relevantes por el usuario, n_{nr} es el número de documentos considerados no relevantes, y α, β y γ son constantes que permiten ajustar el impacto de los documentos relevantes y los no relevantes. Un estudio de sus valores puede encontrarse en (Salton y

Buckley, 1990; Buckley y Allan, 1994; Buckley et al., 1994).

4 Mejora en la eficiencia de la recuperación

Hemos llevado a cabo un pequeño experimento para observar en qué medida la interacción con el usuario, a través de la aplicación de estas técnicas de realimentación puede mejorar los resultados en la Recuperación de Información. Para ello, hemos utilizado una pequeña colección experimental de documentos en español, que ha sido utilizada anteriormente en otros experimentos (Figuerola et al., 2000; Gómez, 1998). La colección se compone de 1074 documentos a texto completo, provenientes del vaciado de parte de las revistas que recibe la biblioteca de nuestra facultad. Es preciso indicar que el sistema no utiliza ningún tipo de descriptor que pudieran dar cuenta del contenido de esos documentos. Se dispone también de 15 consultas para esa colección, así como de una estimación de relevancia efectuada manualmente para cada una de ellas.

Sobre nuestra colección de documentos se aplicó el modelo vectorial tal como se ha descrito en el segundo apartado, obteniéndose la representación vectorizada de los mismos. Se lanzaron y procesaron las preguntas, obteniéndose los resultados ordenados según el índice de similitud para cada una ellas. Seguidamente se examinaron los 15 primeros documentos recuperados, y se marcaron aquellos relevantes para el usuario, así como los no relevantes. En este sentido debemos destacar que este proceso fue llevado a cabo por 32 usuarios diferentes. A continuación se obtuvieron los nuevos pesos para los vectores de las consultas realimentadas, que fueron procesadas de nuevo por el sistema de recuperación, obteniendo nuevos documentos.

Se utilizaron valores para α, β y γ de 1.0, 0.8 y 0.4, respectivamente. Los resultados del proceso se muestran en la figura 1. Las curvas han sido diseñadas de acuerdo con los estándares en evaluación de eficiencia en recuperación de información (Harter y Hert, 1977), y representan la relación entre precisión (*precision*) y exhaustividad (*recall*). La precisión mide la proporción de documentos recuperados que son relevantes. La exhaustividad mide la proporción de documentos relevantes que han sido recuperados. El objetivo es que las dos magnitudes sean lo mayor posible, es decir, recuperar una gran cantidad de documentos relevantes, y al mismo tiempo evitar recuperar documentos no relevantes.

Sin embargo, conseguir este fin es difícil, ya que ambos índices suelen tener un comportamiento inverso, es decir, cuando uno aumenta, el otro tiende a disminuir.

Podemos observar en la figura 1 que la curva precisión-exhaustividad para el caso de consultas realimentadas es francamente mucho mejor que para el caso de las consultas originales. Esto demuestra que la interacción con el usuario a través de la realimentación consigue mejorar los resultados

de forma notoria.

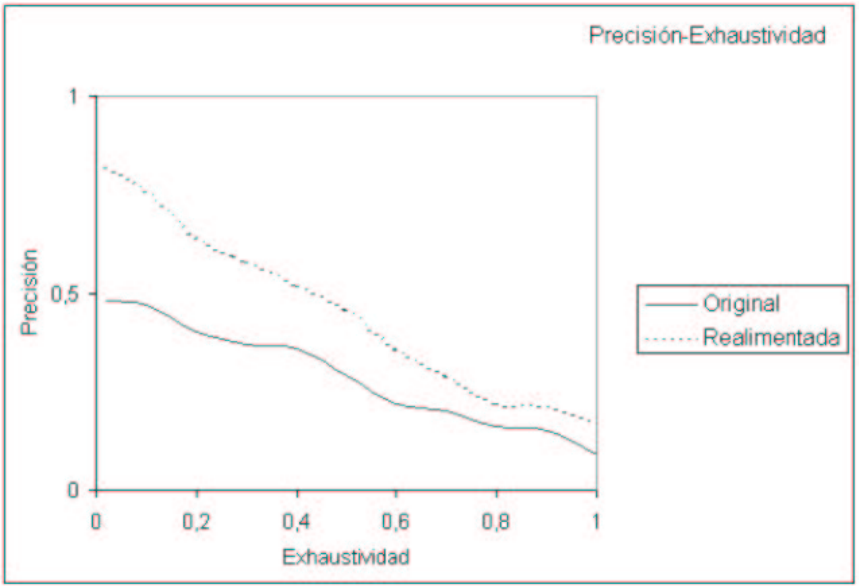

Figura 1: Curvas precisión-exhaustividad

5 Conclusiones

La realimentación de consultas basándose en las estimaciones de relevancia por parte de los usuarios es una técnica que permite que el usuario, tras una primera consulta convencional, refine ésta, proporcionando al sistema ejemplos tanto positivos como negativos acerca de la clase de documentos que desea. Los resultados que se obtienen aplicando esta realimentación mejoran notablemente los resultados iniciales, consiguiendo una mayor cantidad de documentos relevantes recuperados en los primeros lugares, con una reducción del número de documentos no deseados (mayor precisión), así como encontrar un número mayor de documentos relevantes en general (mayor exhaustividad).

Es importante destacar que el proceso es conceptualmente simple y poco costoso en gasto computacional, aspecto muy destacable en sistemas con cientos de miles de documentos.

Un aspecto importante en el proceso de realimentación de consultas es el interfaz de usuario. Debe dotarse al sistema de mecanismos que permitan visualizar los documentos recuperados que desee el usuario, lo cual implica acceso directo al documento, así como la posibilidad de marcar los documentos que considere pertinente y no pertinentes a su necesidad informativa.

Referencias

Belkin, N.J.; Croft, W.B. (1987). Retrieval techniques. // Annual Review of Information Science and Technology, 22, p. 109-145 .

Salton, G.; Buckley, C. (1990). Improving retrieval performance by relevance feedback. // Journal of the American Society for Information Science, 41 (4), 288-297.

Buckley, C.; Allan, J.; Salton, G. (1994). Automatic routing and ad-hoc retrieval using SMART: TREC 2. // Donna Harman, editor, Proceedings of the Second Text Retrieval Conference TREC-2. NIST Special Publication , 1994, 500-215.

Buckley, C.; Salton, G.; Allan, J. (1994). The effect of adding relevance information in a relevance feedback environment. // ACM SIGIR'94, 1994, 292-300.

Figuerola, C.G.; Gómez, R.; López de San Roman, E. (2000). Stemming and n-grams in Spanish: an evaluation of their impact on information retrieval. // Journal of Information Science, 26(6), 461-467.

Gómez Díaz, R. (1998): La recuperación de Información en Español: Evaluación del Efecto de sus Peculiaridades Lingüísticas, Tesina, Universidad de Salamanca, 1998.

Harman, D. (1992a). Ranking Algorithms. //Frakes, W.B.; Baeza-Yates, R. Information retrieval: Data Structures and Algorithms. Prentice-Hall, Englewood Cliffs (NJ), 1992, 363-392.

Harman, D. (1992b). Relevance feedback and other query modification techniques. // Frakes, W.B.; Baeza-Yates, R. Information retrieval: Data Structures and Algorithms. Prentice-Hall, Englewood Cliffs (NJ), 1992, 241-263.

Harter, S.P.; Hert, C.A. (1977). Evaluation of information retrieval systems. // Annual Review of Information Science and Technology, 32, p. 3-94.

Hooper, R. S. (1965). Indexer consistency tests-origin, measurements, results and utilization. Bethesda, MD., 1965.

Rocchio, J.J. (1971). Relevance feedback in Information Retrieval. // Salton, G. (ed.). The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice-Hall, Engelwood Cliffs (NJ), 1991, pp. 313-323.

Salton, G.; Buckley, C. (1988). Term-Witghting Approaches in Automatic Text Retrieval. // Information Processing and Management, 24(5), 513-523.

Salton, G.; McGill, M.J. (1983). Introduction to Modern Information Retrieval. McGraw-Hill, New York, 1983.

Salton, G. (1968). Automatic Information Organization and Retrieval. McGraw-Hill, N.Y., 1968.

Stubbs, E. A.; Mangiaterra, N.E; Martinez, A. M. (2000). Internal quality audit of indexing: a new application of interindexer consytency. // Cataloguing & Classification Quaterly, 28(4), 53-70.