Búsquedas Web con Información de Contexto y Anotaciones Sociales

Gabriel H.Tolosa y Fernando R.A. Bordignon
Universidad Nacional de Luján
Departamento de Ciencias Básicas
{tolosoft, bordi}@unlu.edu.ar

Resumen

Una de las tareas más comunes que realizan los usuarios en la Web es la búsqueda de información utilizando motores de búsqueda tradicionales. Generalmente, éstas se basan en un conjunto de términos que son tratados fuera de contexto alguno. La incorporación de información contextual permite obtener resultados más precisos y puede ser presentada al sistema de formas diferentes. Una fuente posible son los sistemas basados en "anotaciones sociales", los cuales se enriquecen con la participación de los usuarios para organizar información.

En este trabajo se trata el problema de las búsquedas en contexto de literatura científica. En particular, se propone la utilización de un artículo científico como información de contexto para una consulta dada. Además, ésta se complementa con la incorporación de anotaciones sociales provenientes de etiquetas asociadas a los artículos, para ser utilizada en un motor de búsqueda de propósito general existente. Se presenta una propuesta, resultados preliminares y el estado actual de la investigación.

Palabras clave: Búsquedas web, contexto, anotaciones sociales.

1 – Introducción

De todas las tareas que realizan los usuarios en la Web, una de las más comunes es la búsqueda de información utilizando servicios como Google o Yahoo! Search. Generalmente, se realizan búsquedas por palabras clave (*keywords*): el usuario ingresa un conjunto de términos (consulta o *"query"*) con los que "intenta" describir su necesidad de información. La respuesta es un ranking, presentado de acuerdo a aquellos documentos que son más relevantes al *query*.

Generalmente, un *query* es demasiado corto. En un estudio reciente sobre nueve motores de búsqueda, Jansen et al. [7] hallaron que el promedio de términos es menor a 3. En [12] se propone una clasificación de las consultas en cuanto a su precisíon en la especificación o su ambiguedad: a) Query Ambiguo, es aquel que posee más de un significado; b) Query Amplio, es uno que cubre una variedad de subtemas y el usuario está interesado sólo en uno de éstos; y c) Query Claro, tiene un significado específico y cubre un tema acotado. Con esta reducida cantidad de información y si el usuario propone una consulta tipo "a" o "b", la tarea resulta dificultosa, por lo que la incorporación de otros "indicios" acerca de la necesidad de información es de utilidad.

Los motores de búsqueda de propósito general – de forma normal – tratan a la consulta de forma aislada a su contexto. Cuando retornan los resultados, algunos de éstos pueden ser de utilidad, pero esta discriminación depende del contexto del *query* [8], esto es, información extra que pueden ayudar a determinar su correcta interpretación. Aquí hay que diferenciar el uso del contexto

de una consulta con personalización de las búsquedas. En la última, se hace uso de información acerca del usuario para proveer un resultado sesgado hacia sus preferencias, intereses, gustos, conocimiento, ubicación geográfica, y demás; lo que le permita mantener una relación más "individualizada" con el sistema de búsquedas [4]. Existen algunos servicios de búsquedas que hacen uso de información de contexto y/o personalización, ya sea implícita o explícita por parte de los usuarios. No obstante, éstos aún son experimentales y no está clara áun su eficiencia.

Una de las fuentes para obtener información de contexto implícita son los sistemas basados en "anotaciones sociales" (*social tagging*) los cuales se enriquecen con la participación de los usuarios – de forma distribuida – para organizar información. Esta tarea se realiza habitualmente mediante servicios de *bookmarking* y permiten agregar a los elementos de información una serie de etiquetas descriptivas (*tags*) que los usuarios consideran que pueden ser útiles al momento de querer recuperar el objeto en cuestión. Los primeros servicios de este tipo – como Del.icio.us[1] o Digg[2] – permiten organizar direcciones electrónicas (URLs); luego aparecieron sistemas específicos como CiteULike[3] o Bibsonomy[4] orientados a la comunidad académica, que permiten organizar artículos científicos, journals y libros de diversos temas. Toda esta nueva información creada a partir de la colaboración de los usuarios es una fuente valiosa para incorporar al problema de las búsquedas [5].

La utilización de información de contexto en búsquedas web fue estudiado por Lawrence [8], quien identificó la necesidad de sistemas que incorporen este elemento – junto con esquemas de personalización – para ayudar a los usuarios en su tarea. Kraft et al. [6] propusieron tres algoritmos para aumentar la relevancia de los resultados. En particular, hallaron que la técnica basada en la reescritura de la consulta original generaba un aumento de la eficiencia. En [2] se estudió el problema de integrar anotaciones sociales con las búsquedas web y propusieron dos algoritmos (SocialSimRank y SocialPageRank) para mejorar la calidad de las respuestas.

En este trabajo se trata el problema de las búsquedas en contexto de literatura científica. Se propone la utilización de un artículo científico como información de contexto para una consulta, complementada con la incorporación de anotaciones sociales. En este modelo se considera que una consulta es una tupla <Q, d>, donde Q es el conjunto de los términos de la consulta y d es un artículo científico (*paper*). El objetivo es contruir un "nuevo" *query*, definido como <Q`, K>, donde $Q$` es el conjunto de términos de la consulta "ajustados" al contexto utilizando d, y K es un conjunto de palabras clave obtenidas de un sistema social como los mencionados. Esta nueva consulta será enviada a un motor de búsquedas tradicional. Se espera obtener más cantidad de documentos relevantes en las primeras posiciones de la lista de respuesta a partir de un *query* más preciso.

2 – Descripción de la Propuesta

En este trabajo se propone la redefinición de una consulta a partir de información de contexto y anotaciones sociales provenientes del sistema de *social bookmarking* CiteULike, orientado a la recuperación de literatura científica. Este sistema ofrece un servicio gratuito con interface web que permite organizar y compartir artículos científicos, permitiendo su organización mediante etiquetas propias (*folksonomías* [10]). Si bien las etiquetas que agregan los usuarios están asociadas a un artículo, también es posible extraer relaciones entre éstas y las *keywords* del artículo, sus autores y otros descriptores. La idea es enviar la "nueva" consulta a motores de búsqueda tradicionales a través de su interface web natural.

1 http://del.icio.us/
2 http://digg.com/
3 http://www.citeulike.org/
4 http://www.bibsonomy.org/

De forma nomal, una consulta Q es un conjunto de términos $t_1...t_n$; habitualmente sin repeticiones. Sobre esto último se destaca que son "sin repeticiones" solamente por uso y no por diseño, ya que algunos motores de búsqueda ponderan los términos de la consulta y retornan listas de respuesta diferentes[5]. En este modelo se redefine la consulta como una tupla <Q, d>, donde *Q* es el conjunto de los términos de la consulta y *d* es un artículo científico (*paper*) entregado por el usuario y que corresponde al contexto. Esta información es la entrada (*input*) de una función que retorne un "nuevo" *query*, definido como <Q`, K>, donde *Q`* es el conjunto de términos de la consulta "ajustados" al contexto utilizando *d* y *K* es un conjunto de palabras clave $k_1...k_n$; obtenido a partir del procesamiento de las relaciones entre los documentos de CiteULike. Los parámetros Q y d de la consulta son provistos por el usuario, mientras que K proviene del sistema.

La motivación detrás de esta propuesta está dada por la problemática que se presenta cuando un usuario (en este caso particular, alguien buscando literatura científica) desea encontrar documentos relacionados o "similares" en cuanto a temática respecto de uno que conoce y posee. El problema – entonces – consiste en poder extraer del documento d y de la información del sistema de anotaciones sociales un nuevo conjunto de términos que permitan delimitar el contexto de la consulta para construir un nuevo "*query*". Por ejemplo, si Q = "evaluación búsquedas web" y d es el artículo "Evaluación de Buscadores Web", el nuevo query Q' podría ser "evaluación búsquedas web precisión cobertura".

El enfoque considera la extracción de información del artículo [9] desde partes estructurales precisas: resumen, introducción, conclusiones, GIST [13] o combinaciones de éstas. Luego, se incorporarán relaciones extraidas de CiteULike. Para ello, se propone la construcción del grafo de relaciones G = (V, A), donde los V es el conjunto de vértices que corresponden a los documentos en el sistema y A, el conjunto de aristas que conectan dos documentos. Se define que existe una arista entre v_i y v_j si y solo si contienen un número mínimo de *tags* comunes. Luego, a partir de la información de dichos documentos (título, resumen, etc.), se retroalimenta el modelo del "nuevo" *query*. Para ello se define la función que evalua – dados los parámetros mencionados – cuáles características "representan" de mejor manera el contexto de la consulta. Aquí se consideran medidas de similitud textual (TF/IDF [1], BM25 [11]) entre Q, d y K, y medidas de análisis de grafos para las relaciones extraidas de CiteULike.

3 – Resultados Iniciales

En esta sección se presentan los resultados iniciales provenientes del estudio de CiteULike como fuente de anotaciones sociales. De estos resultados se derivarán las características que mejor aporten a la redefinición del *query*.

Se trabajó con un *snapshot* oficial de CiteULike del 13 de febrero de 2008. Los datos básicos se presentan en la tabla 1. Un primer dato es que el promedio de tags por documento es 2.8, y es similar al promedio de términos en las consultas reportado en [7]. Esto podría sugerir que los usuarios anotan en los documentos aquellos términos con los que lo buscarían, es decir, el conjunto de las etiquetas de un documento (para un usuario) sería similar a un *query*. La riqueza en el sistema radica en que múltiples usuarios anotan un mismo documento con etiquetas diferentes. No obstante, éstas deben ser cuidadosamente analizadas para determinar si corresponden al contexto del *query* y del artículo ejemplo.

5 Puede probar enviando – por ejemplo – a Google la consulta "mercedes" y luego "mercedes mercedes" y comparar las listas de resultados.

Cantidad Total de Documentos	824.629	Cantidad Total de Etiquetas (*Tags*)	2.371.986
Cantidad Total de Usuarios	22.392	Cantidad Total de Etiquetas Únicas	144.666

Tabla 1 – Datos básicos del *snapshot* estudiado.

3. 1 – Frecuencia de las Etiquetas

Se analizó la frecuencia de aparición de las etiquetas y se conformó un ranking. Se comprobó que éste sigue una ley de Zipf [14], la cual establece que frecuencia de la i-ésima palabra es $1 / i^{\Theta}$ veces que la más frecuente (es una *power-law*). Esta es una ley clásica en el análisis de textos en lenguaje natural, que propone que los autores tienden a escribir utilizando mayormente términos conocidos (ley del menor esfuerzo). Esta característica no necesariamente debe aparecer en sistemas de anotaciones, no obstante, esto puede estar relacionado con el hecho que el sistema permite seleccionar entre las etiquetas que un usuario ya utilizó al ingresar un nuevo documento.

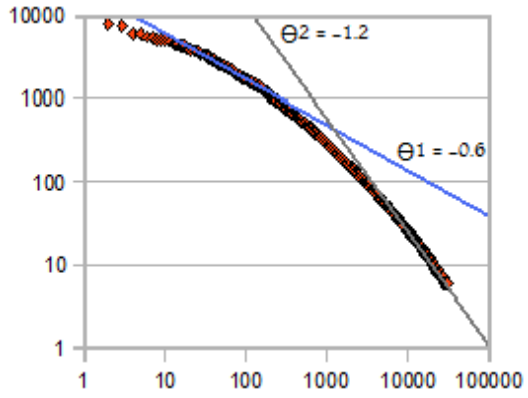

Figura 1 – Distribución de la Frecuencia de las Etiquetas y sus dos rectas de ajuste. Los ejes-xy están en log-log.

1	bibtex-import	17898
2	no-tag	13234
3	review	8135
4	evolution	7476
5	research	6011
6	support	5994
7	learning	5537
8	govt	5483
9	analysis	5186
10	animals	5136
11	model	5098
12	theory	5039
13	human	4953
14	humans	4503
15	models	4497

Tabla 1 – Etiquetas más frecuentes

El ajuste de la ley de Zipf se realizó mediante dos curvas, una para los valores hasta la posición 1000 del ranking, las cuales acumulan el 49% de las etiquetas totales y otra para los superiores. En el primer caso, se halló un valor $\Theta_1 = -0.6$, mientras que en el segundo, $\Theta_2 = -1.2$. En el caso de las etiquetas más frecuentes se hallaron algunas que no aportan a la descripción del documento sino que hablan del tipo o de la forma de ingreso al sistema (por ejemplo, la primera etiqueta: bibtex-import). Las 15 etiquetas más frecuentes se presentan en la tabla 1. Esto sugiere que hay que estudiar un umbral a partir del cual se consideran etiquetas "útiles".

3. 2 – Documentos y Etiquetas

Se estudió la distribución de etiquetas por documentos y se encontró que el 46% de éstos solo poseen 1, el 40% tiene entre 2 y 5 y el 14% posee 6 o más etiquetas. Estos valores son similares tanto para las cantidades de etiquetas únicas o totales. Siendo el promedio de etiquetas por documentos ~ 3, es interesante considerar inicialmente el conjunto intermedio (entre 2 y 5 *tags*). Los porcentajes de este grupo resultaron: 15, 12, 8 y 5% para 2, 3, 4 y 5 etiquetas respectivamente.

4 – Contexto del Trabajo y Discusión

Este trabajo se encuentra en el marco del proyecto de investigación "*Modelos y Servicios de Información sobre Sistemas Complejos en Espacios Académicos y Científicos*", aprobado por el Departamento de Ciencias Básicas de la Universidad Nacional de Luján. Dicho proyecto tiene entre

sus objetivos diseñar estrategias para aumentar la eficiencia de los servicios de búsqueda en redes globales, en particular en la Web.

En este artículo se presenta una línea de investigación del proyecto, con una propuesta para realizar búsquedas con información de contexto y anotaciones sociales; y algunos resultados preliminares. Se está diseñando un esquema de identificación de contexto basado en métricas de similitud de texto y se construyó un grafo con los documentos de CiteULike de acuerdo a lo propuesto. Por otro lado, se deberán realizar experimentos de evaluación que permitan definir las características a utilizar en la función que integre las fuentes y genere Q` y el conjunto K de téminos de contexto. Para la evaluación, se considera adaptar la propuesta realizada por Chowdhury [3].

5 – Referencias

[1] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, 1999.

[2] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei and Z. Su. Optimizing web search using social annotations. Proceedings of the 16th international conference on World Wide Web, 2007.

[3] A. Chowdhury and I. Soboroff. Automatic Evaluation of World Wide Web Search Services. Proceedings of SIGIR'02, 421 – 422, 2002.

[4] Z. Dou, R. Song and J. Wen. A large-scale evaluation and analysis of personalized search strategies. Proceedings of the 16th international conference on World Wide Web, 2007.

[5] M. Kipp. Tagging Practices on Research Oriented Social Bookmarking Sites. Proceeding of Information Sharing in a Fragmented World: Crossing Boundaries Conference. CAIS/ACSI, 2007.

[6] R. Kraft, C. Chang, F. Maghoul, R. Kumar. Searching with context. In WWW'06. Proceedings of the 15th international conference on World Wide Web, 2006.

[7] B.J. Jansen and A. Spink. How are we searching the world wide web?: A comparison of nine search engine transaction logs. Information Processing and Management, 42(1), 2006.

[8] S. Lawrence. Context in Web Search, IEEE Data Engineering Bulletin, Vol.23, Nº 3, 2000.

[9] P. Lavallén, F. Bordignon y G. Tolosa. Reconocimiento Automático de Artículos Científicos. VII Workshop de Investigadores en Ciencias de la Computación. WICC 2005.

[10] A. Mathes. Folksonomies – Cooperative Clasification and Communication through Shared Metadata. 2004. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

[11] S. E. Robertson, H. Zaragoza, and M, Taylor. Simple BM25 extension to multiple weighted fields. Proceedings of the 13ª ACM Conf. on Information and Knowledge Management, CIKM '04.

[12] R. Song, Z. Luo, J. Wen, Y. Yu, H. Hon. Identifying Ambiguous Queries in Web Search. Proceedings of the 16th international conference on World Wide Web, 2007.

[13] G. Tolosa, J. Peri y F. Bordignon. Experimentos con Métodos de Extracción de la Idea Principal de un Texto sobre una Colección de Noticias Periodísticas en Español. XI CACIC, 2005.

[14] Zipf, G. Human Behaviour and the Principle of Least effort. Addison-Wesley, 1949.