Selección de Recursos Distribuidos en Ambientes Dinámicos Basados en Web

Santiago Banchero[1], Fernando R. A. Bordignon, Gabriel H. Tolosa
{sbanchero, bordi, tolosoft}@unlu.edu.ar

Resumen

La masificación de las comunicaciones de datos y el surgimiento de múltiples fuentes de información en-línea ha generado le necesidad de poner atención en el problema de realizar búsquedas sobre repositorios que se encuentran distribuidos. Este problema puede dividirse en tres partes: la representación de cada fuente a los efectos de permitir las búsquedas, la selección de las adecuadas de acuerdo a una consulta y la fusión de los resultados para presentar al usuario.

Este artículo presenta los primeros avances en el trabajo de construcción de descripciones de recursos distribuidos y evaluación de algoritmos de selección. El objetivo es integrar y adaptar distintos algoritmos pertenecientes al área de Recuperación de Información Distribuida para que funcionen conjuntamente con fuentes de información heterogéneas en ambientes dinámicos basados en Web. Se utilizarán recursos que presten servicio de sindicación de contenido y así poder evaluar cómo responden los algoritmos de selección de recursos distribuidos en espacios acotados como son blogs y otras fuentes que utilizan esta modalidad de publicación de contenidos.

Palabras clave: Recuperación de información distribuida, representación y selección de recursos, sindicación de contenidos.

1 – Introducción

El área de Recuperación de Información ha sido pionera en la tarea de buscar y rankear documentos relevantes a partir de una necesidad de información del usuario [1,2], operando sobre grandes volúmenes de información, generalmente en documentos de texto y tradicionalmente bajo esquemas centralizados.

Con el acentuado crecimiento de las comunicaciones y la expansión de Internet como plataforma de intercambio de información surge la necesidad de integrar distintas fuentes. Nace así la Recuperación de Información Distribuida RID (también llamada *Federated Search*) [1] que pretende dar respuestas al problema de la Recuperación de Información en un nuevo ambiente ampliado por el crecimiento de repositorios, tanto dentro de las organizaciones (Figura 1a) como de la red global de información (Figura 1b). La RID tiene como objetivo principal desarrollar modelos y estrategias para obtener el mayor beneficio de estas fuentes distribuidas para responder a las distintas necesidades de información de sus usuarios, que perciben al sistema como único, independientemente del número de fuentes que existan. El proceso es totalmente transparente al usuario por lo tanto no percibe la complejidad del mismo.

La RID incluye tres subproblemas a estudiar [1, 3, 6]: a) DESCRIPCIÓN DE LOS RECURSOS, es decir, cómo se representa la información que se encuentra distribuida en repositorios, denominados corpus de documentos o bases de datos textuales, b) la SELECCIÓN DE RECURSOS, donde a partir de una necesidad de información y un conjunto de descripciones de éstos se decide cuáles serán los que tengan mayor probabilidad de satisfacer la consulta. Por último, c) la FUSIÓN DE LOS RESULTADOS consiste en la integración de los resultados retornados por las consultas a n bases de datos formando una única lista que presenta el ranking de resultados.

---

[1] Actualmente desarrollando su trabajo final de licenciatura en esta temática.

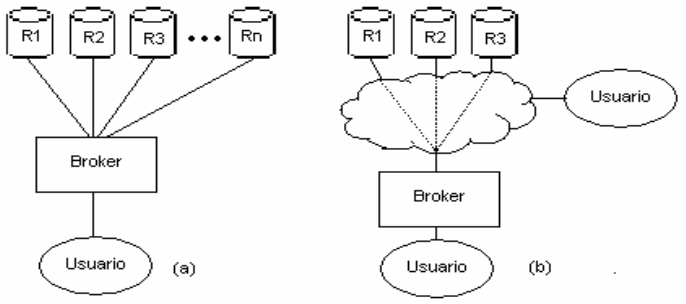


Figura 1

El objetivo los un sistema de RID es proveer una interfaz sencilla para poder acceder a todos los recursos distribuidos, orientar cada necesidad de información del usuario hacia el recurso que mejor la satisfaga y luego fusione lo seleccionado en una única lista de resultados [10]. Un caso clásico en la Web son los metabuscadores (Metasearch[2], Mamma[3], etc). Como se ha mencionado estos sistemas tratan con la problemática fundamental de la RID [1]. La característica principal que justifica el uso de sistemas de RID es que permite realizar búsquedas más exhaustivas que los buscadores tradicionales. También reducen el tráfico en la red. Esto es posible gracias a que evita tener que realizar una consulta a todos los repositorios ya que se puede contar con una descripción mínima de este y así saber si esta en condiciones de satisfacer esa necesidad de información. De esta manera solo se consultaran aquellas que estén en condiciones de responder.

Los trabajos realizados hasta la actualidad, y los más interesantes, incluyen algoritmos para la construcción de descripciones de recursos en ambientes no cooperativos como Query Based Sampling [15], Capture-Recapture, etc. Estos algoritmos permiten obtener algunos valores que no están disponibles, por las características del entorno. Por ejemplo, el tamaño de la colección en cantidad de documentos y otros datos estadísticos de estos repositorios que son requeridos a la hora de seleccionar.

Además, los algoritmos de selección de recursos [3, 9] que utilizan modelos de espacios vectoriales de IR tradicional como gGLOSS [14], redes de inferencia bayesianas como INQUERY [12], CORI [5], ReDDE [10]. Estos últimos son los más utilizados por la comunidad de investigadores de RID. El objetivo fundamental de estos algoritmos es retornar un pequeño conjunto de bases de datos (recursos) que contengan la mayor cantidad de documentos relevantes para una consulta. Los sistemas de RID son también una solución a los problemas de escalabilidad que presentan los motores de búsqueda tradicionales que deben manejar grandes volúmenes de información y utilizar demasiados recursos de hardware, software y ancho de banda. La solución aportada por la RID es más robusta, facilita el mantenimiento de índices ya que no utilizan un único índice central sino que se propone la utilización de un índice por cada recurso [16].

Por otro lado, existe todo un nuevo espacio de publicación que puede ser accesible a través de servicio de sindicación de contenido, que permite trabajar de manera opuesta a la idea original de publicar en un sitio web que los usuarios deban obligatoriamente visitar [13] por ejemplo, diarios como Clarín, La Nación entre otros. También existen buscadores verticales que operan sobre espacios acotados como: TECHNORATI, GOOGLE BLOG SEARCH[4], FEEDSTER[5].

---

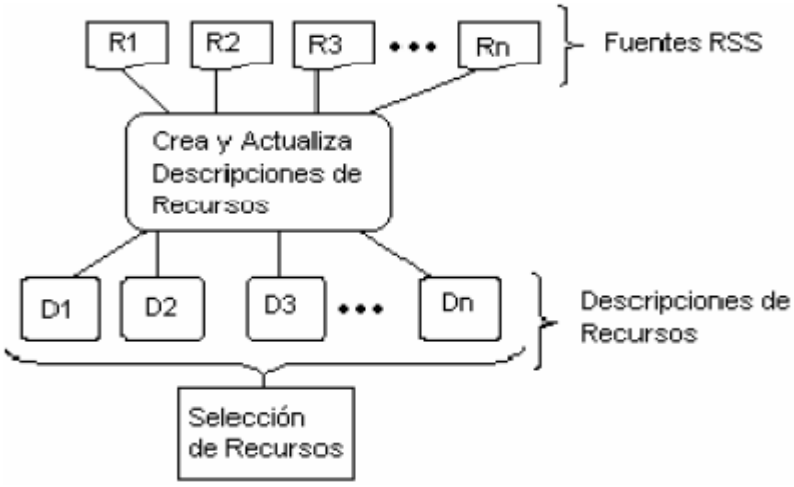
2 http://metasearch.com
3 http://www.mamma.com
4 http:// blogsearch.google.com
5 http://www.feedster.com

En este trabajo de investigación se propone la integración de técnicas de RID y las nuevas formas de publicación de contenido. Es decir, la utilización de recursos Web existentes que presten servicio de feeds – RSS, ATOM y RDF – para publicar su contenido. Principalmente se apunta a Blogs y sitios Web personales de iguales características a la hora de publicar contenido. La utilización de estas tecnologías para construir descripciones de recursos constituye una alternativa novedosa ya que conduce a un ambiente de trabajo híbrido donde no hay una cooperación absoluta por parte de la fuente sino que es parcial, y se realiza a través de feeds. Tanto la sindicación de feeds como los Blogs son parte de la evolución de la Web y forman parte de un movimiento conocido como Web 2.0 [4] cuya filosofía es la reutilización de herramientas Web existentes y el aprovechamiento de la inteligencia colectiva.

2 – Objetivos de la Propuesta

El objetivo de este trabajo es integrar y adaptar distintos algoritmos pertenecientes al área de RID para que funcionen conjuntamente con fuentes de información heterogéneas en Ambientes Dinámicos Basados en Web. Se trabajará con algoritmos que corresponden al primero y segundo subproblema de RID, descripción de los recursos y selección de recursos, respectivamente. Se requiere: a) una aplicación que recupere y almacene documentos XML en formatos de *feeds* RSS, ATOM y RDF. b) Definir una estructura de datos propia para almacenar el contenido de cada una de las fuentes. c) Adaptar un algoritmo de selección de recursos para tratar con el tipo de objeto recuperados.

Puntualmente, se propone un modelo de BD textual adaptando un algoritmo de selección de recursos basado en los clásicos como CORI [5] y ReDDE [10]. Se desarrollará una herramienta que permita generar descripciones de recursos de fuentes heterogéneas en español y que permita el estudio de la evolución del lenguaje de cada fuente. Para poder realizar esto la aplicación deberá recuperar documentos publicados a través de sindicación de documentos. A continuación se incluye



un gráfico que ilustra la arquitectura del modelo:

3 – Resultados Preliminares

Esta investigación surge como la evolución del trabajo realizado en el marco de un curso de Inteligencia Artificial en la Lic. en Sistemas de Información de la UNLu denominado “Algoritmo de Selección de Recursos Basado en Redes de Creencia para Recuperación de Información Distribuida” [17].

El algoritmo utiliza las representaciones de las base de datos textuales y de las consultas para generar evidencia acerca de la capacidad de cada una de las fuentes para satisfacer cada requerimiento. La idea inicial de esta propuesta surge del algoritmo de selección de recursos CORI, el cual utiliza una red de inferencia Bayesiana y una versión adaptada de la fórmula Okapi de normalización de frecuencias de términos. En la nueva aproximación se modifica el diseño de la red, las asignaciones de pesos a cada uno de los términos y la fórmula de ponderación de cada base de datos.

En las pruebas iniciales utilizando el algoritmo propuesto obtuvo una performance comparable con el modelo CORI (Gráfico 1 y 2), utilizando las colecciones de prueba CISI y CACM de acuerdo a la metodologia TREC, la cual ha sido estudiada previamente [18]. En dichos experimentos, se utilizaron 3 métodos de ponderación de términos: frecuencia relativa (PFR), Estimador de máxima Especificidad (PMNi) y suavizado del MLE (PSM).

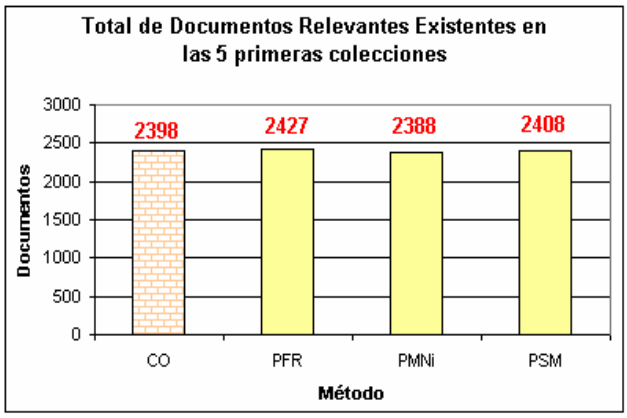


Gráfico 1

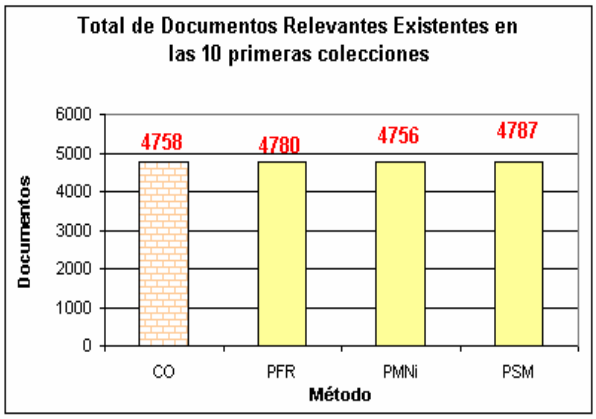


Gráfico 2

Actualmente, se está trabajando con la construcción de bases de datos textuales a partir de los recursos RSS recuperados a los efectos de caracterizar cada base de dato textual y poder retroalimentar el modelo de selección al ambiente en cuestión.

4 – Referencias

[1] Callan J. Distributed Information Retrieval. In W.B. Croft, editor, Advances in information retrieval, chapter 5, pages 127-150. Kluwer Academic Publishers, 2000.

[2] Si, L., & Callan, J., Modeling search engine effectiveness for federated search, Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, August 15-19, 2005, Salvador, Brazil

[3] French J. C. , Powell A. P., Callan J., Viles C. L., Emmitt T., Prey K. J., Mou Y., Comparing the performance of database selection algorithms, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, p.238-245, Berkeley, California, United States (August 15-19, 1999)

[4] O' Reilly T. Presidente y CEO de O' Reilly Media, INC. Qué es web 2.0. Patrones del diseño y modelos del negocio para la siguiente generación del software.

[5] Callan, J, Lu Z., Croft W. B. Searching distributed collections with inference networks, Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, p.21-28, Seattle, Washington, United States (July 09-13, 1995)

[6] French J. C, Powell A. L., Viles C. L., Emmitt T., Prey K. J. Evaluating database selection techniques: a testbed and experiment, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p.121-129, August 24-28, Melbourne, Australia (1998)

[7] Xu J., Callan, J, Effective retrieval with distributed collections, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p.112-120, Melbourne, Australia (August 24-28, 1998)

[8] Si, L., & Callan, J., Unified utility maximization framework for resource selection, Proceedings of the thirteenth ACM international conference on Information and knowledge management, November 08-13, 2004, Washington, D.C., USA

[9] Si, L., & Callan, J. Relevant document distribution estimation method for resource selection, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, Toronto, Canada (July 28-August 01, 2003)

[10] Si, L., & Callan, J. Distributed information retrieval with skewed database size distributions. In Proceedings of the national conference on digital government research. (2003a)

[11] Shokouhi M., Zobel J., Scholer F., and Tahaghoghi S. M. M. Capturing collection size for distributed non-cooperative retrieval. In SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 316–323, New York, NY, USA, 2006. ACM Press.

[12] Callan J. P., Croft W. B., and Harding S. M. The INQUERY retrieval system. In Proceedings of the Third International Conference on Database and Expert Systems Applications, pages 78{83, Valencia, Spain,. Springer-Verlag (1992)

[13] Hammond T., Hannay T., and Ben Lund. The Role of RSS in Science Publishing. Syndication and Annotation on the Web.D-Lib Magazine. Volume 10 Number 12. ISSN 1082-9873. (December 2004)

[14] Gravano L. and Garcia-Molina H. Generalizing GlOSS to vector-space databases and broker hierarchies. Technical Report STAN-CS-TN-95-21, Stanford University. Available as ftp: //db. Stanford.edu/pub/gravano/-1995/stan.cs.tn.95.21.ps (May 1995)

[15] Callan, J. and Connell, M. Query-based sampling of text databases. Technical Report IR-180, Center for Intelligent Information Retrieval, Department of Computer Science, University of Massachusetts. (1999)

[16] Baeza-Yates R., Castillo C., Junqueira F, Plachouras V. and Silvestri F. Challenges in Distributed Information Retrieval (invited paper). In ICDE (Istanbul, Turkey). (April 2007)

[17] Banchero, S.; Tolosa, G. H.; García, O. y Bordignon, F. R.A. Algoritmo de Seleccion de Recursos Basado en Redes de Creencia para Recuperación de Información Distribuida. UNLu Departamento de Ciencias Básicas Laboratorio de Redes.(2005)

[18] Tolosa, G..; Bordignon, F. R. A., Peri, J. A., Banchero S. Creación de una colección de prueba de literatura científica en español para evaluar sistemas de recuperación de información. UNLu Departamento de Ciencias Básicas (WICC 2005)