Recuperación de Información Distribuida sobre Bases de Datos Textuales Basadas en Sindicación

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

Resumen

El área de investigación en Recuperación de Información Distribuida se ha desarrollado sostenidamente en los últimos años, esto es consecuencia de la creciente expansión de repositorios de información textual y el furor de aplicaciones basadas en tecnologías Web 2.0, como – por ejemplo – blogs, periódicos digitales y wikis.

En este trabajo se aborda el tema de caracterización de bases de datos textuales distribuidas, en particular, aquellas basadas en sindicación de contenidos. Se presentan los resultados preliminares de esta tarea donde una de las principales dificultades radica en como tratar de forma eficiente fuentes de información heterogéneas, distribuidas y generadas en forma dinámica.

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

1 – Introducción

La Recuperación de Información Distribuida [Callan, 2000] es un área de investigación que se ha desarrollado sostenidamente en los últimos años a partir de la expansión de los repositorios de información textual en diferentes organizaciones (empresas, universidades, etc.) y la aparición y rápido desarrollo de nuevos servicios de información en Internet, como – por ejemplo – blogs, periódicos digitales y wikis.

Algunos de estos servicios generan información con alta frecuencia (días, horas) por lo que las colecciones de documentos que poseen tienen un alto dinamismo, en particular, en cuanto su crecimiento. Uno de los mecanismos existentes para la distribución de nueva información hacia los usuarios en la sindicación de contenidos, la cual permite trabajar de manera opuesta a la idea original de publicar en un sitio web que los usuarios deban obligatoriamente visitar [Hammond et al., 2004]. Los sistemas de sindicación de contenidos se basan en la idea de publicar su contenido (o un resumen de éste) en un formato específico en un archivo XML (normalmente utilizando los protocolos RSS, ATOM y RDF). Dicho archivo contiene múltiples elementos de información denominado *feeds* y puede ser continuamente recuperado por un software lector por parte de los usuarios interesados en tal tema.

La información publicada mediante este mecanismo posee algunas características particulares. En general, son elementos de información de poca longitud que incorporan algunos otros atributos como URL del recurso completo, fecha y hora de publicación, autor y demás. Estas características particulares generan que una aplicación de recuperación de información tradicional no necesariamente se adapte adecuadamente y – además – el factor temporal resulta particularmente importante de tratar en sistemas de búsquedas.

Debido a la naturaleza distribuida de las fuentes de información, un enfoque basado en técnicas del área de Recuperación de Información Distribuida [Callan, 2000] resulta adecuado,

principalmente teniendo en cuenta las características mencionadas. En esta área se tratan principalmente 3 subproblemas [Callan, 2000; French et al., 1999; French et al., 1998]:

a) DESCRIPCIÓN DE LOS RECURSOS: representar la información que se encuentra distribuida en repositorios de manera de poder caracterizar cada una.

b) SELECCIÓN DE RECURSOS: A partir de una necesidad de información y un conjunto de descripciones de recursos de debe decidir los adecuados, es decir, aquellos que tengan mayor probabilidad de satisfacer la consulta.

c) FUSIÓN DE RESULTADOS: Luego de las consultas realizadas, se deben integrar los resultados retornados a las n bases de datos para armar una única lista de resultados (ranking) para presentar al usuario.

Una aplicación eficaz de modelos de Recuperación de Información Distribuida requiere de adaptar primero las técnicas de descripción de recursos a los elementos de información en cuestión teniendo en cuenta sus características. A partir de la efectiva caracterización de los elementos de información objeto de estudio se estudiará si las técnicas de selección de recursos existentes son adecuadas o requieren de modificaciones y/o extensiones.

Este trabajo corresponde a la propuesta presentada en [Banchero et al., 2007] en la que se propone un modelo de BD textual para *feeds* y un algoritmo de selección de recursos "ad-hoc". Este último, basado en los clásicos como CORI [Callan, 1995] y ReDDE [Si & Callan, 2003], pero incorporando las características propias de los elementos de información en estudio. En este artículo se presenta la caracterización de las bases de datos textuales formadas por la recuperación mediante sindicación. Particularmente, se estudian las propiedades clásicas de los modelos de bases de datos estáticas como vocabulario, frecuencia de palabras y tamaños de los documentos. Como extensión de éstos para bases de datos dinámicas, se tienen en cuenta las distribuciones de los valores de las propiedades en función del tiempo.

2 – Modelos de Documentos

Los modelos para documentos en bases de datos textuales son los clásicos para el texto en lenguaje natural [Baeza & Ribeiro, 1999] y están basado en evidencia empírica. Éstos incluyen el estudio de las palabras en el vocabulario (modelado mediante la ley de Heaps), la distribución de frecuencias de las palabras (Ley de Zipf), distribución de los tamaños de los documentos y de las palabras dentro de éstos. Se han realizado estudios sobre bases de datos estáticas y también sobre la Web.

2.1 – Distribución de las Palabras

Uno de los modelos clásicos corresponde a la distribución de las frecuencias de las palabras en el texto. La ley de Zipf [Zipf, 1949] establece que la frecuencia de la i-ésima palabra es $1 / i^{\sigma}$ veces que la más frecuente. Una distribución de Zipf se caracteriza por tener pocas observaciones muy frecuentes y muchas poco frecuente. Al graficarse en escala log-log se presenta un recta. A modo de ejemplo, en la figura 1 se muestra la distribución de los términos hallados en Wikipedia en Noviembre de 2006, junto con varias rectas de ajuste.

El valor del parámetro σ depende del texto y del idioma [Gelbukh, 2001]. Por ejemplo, para ajustar a datos reales de textos en inglés, el valor se encuentra entre 1.7 y 2.0. En el estudio de Baeza-Yates para una muestra de la web, se encontró $\sigma = 1.59$, un valor más pequeño que el hallado para texto en ingles.

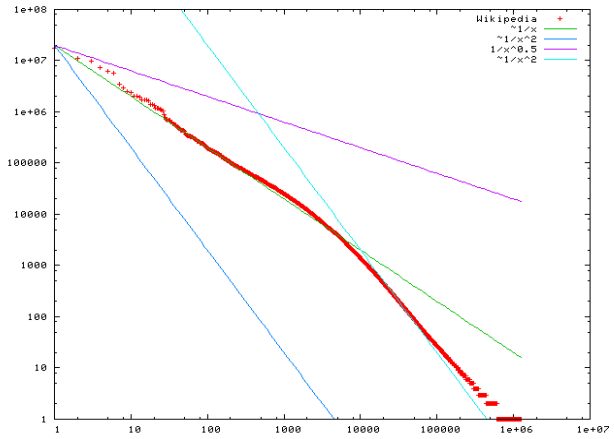

Figura 1 - Distribución de los términos en Wikipedia, Nov-2006. Ejes x,y en escala log-log. [Fuente: http://en.wikipedia.org/wiki/Zipf's_law]

2.2 – Tamaño del Vocabulario

De manera similar a la ley de Zipf, existe otra ley empírica que describe el comportamiento de los términos dentro de un texto escrito denominada ley de Heaps [Heaps, 1978]. La misma permite predecir el crecimiento del vocabulario (palabras únicas) en relación con el tamaño del texto (cantidad de palabras totales). En particular, postula que el tamaño del vocabulario (y su crecimiento) es una función del tamaño del texto, de la forma:

$$V = K \cdot N^{\beta}$$

donde:

N: Es el tamaño del documento (cantidad de palabras)

K: Constante que depende del texto, típicamente entre 10 y 100.

ß: También es una constante que depende del texto, donde $0 < ß < 1$, aunque sus valores típicos se encuentran entre 0.4 y 0.6.

Por ejemplo, para textos en inglés: $10 < K < 20$ y $0.5 < ß < 0.6$

3 – Resultados Preliminares

Para los experimentos preliminares se seleccionó un subconjunto de las fuentes y se las dividió en tres categorías. La producción de *post* por fuente no es homogénea y la intuición detrás de esta decisión es que el ámbito del productor es una de las variables que lo determinan. Por ejemplo, una fuente de un periódico produce un número de *posts* diarios, mientras que un usuario particular puede hacerlo de forma más heterogénea. En la tabla 1 se presentan algunos datos sobre las fuentes estudiadas y las categorías formadas.

Se estudiaron las distribuciones de frecuencias de términos. Inicialmente, se calculó el ajuste a la ley de Zipf para cada categoría. Aquí se supone que al escribir – normalmente – los autores suelen preferir palabras más conocidas sobre aquellas menos conocidas. Si el valor del parámetro σ de la distribución es mas grande, entonces ésta es más sesgada y – por ende – existen un menor uso de la riqueza de la lengua. Los resultados se presentan en la figura 1.

Otro estudio que se realizó es el de crecimiento del vocabulario. Para esto, se calculó el ajuste a la ley de Heaps para cada categoría, figura 2. Aquí se supone que a medida que la cantidad

de post se incrementa el vocabulario se va enriqueciendo. La curva de ajuste de Heaps se consiguió con un K = 30 y un β = 0.82.

Fuentes	URL	Categorias	Cantidad de *Posts*	Cantidad de Términos	*Post*/Día
1	http://alt1040.com/	Personal	2385	69475	7,8
2	http://www.ojobuscador.com/	Empresa	1780	53826	6,1
3	http://barrapunto.com/	Empresa	1541	71913	5,2
4	http://www.milenio.com/	Diario	1355	50209	4,5
5	http://www.javahispano.org/	Empresa	797	76230	2,8
6	http://sociedadenred.com.ar/	Personal	326	9145	1,2
7	http://www.agendaiquique.com/	Empresa	325	9634	1,9
8	http://beatrizgarrido.nireblog.com/	Personal	273	114908	2,3
9	http://mnm.uib.es/	Personal	151	4959	0,9
10	http://www.clarin.com/ **(*)**	Diario	128	64800	1,1

(*) La fuente número 10 corresponde únicamente al canal "El Mundo" del diario Clarín.

Tabla 1 – Fuentes estudiadas en los experimentos preliminares

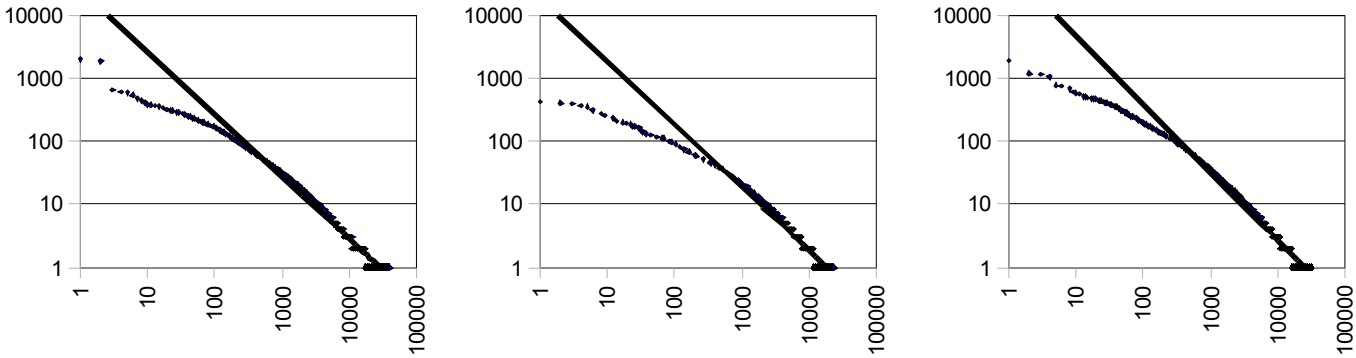

Figura 1 – Distribuciones de frecuencias por categoría con su recta de ajuste a la ley de Zipf. El eje *x* es el ranking, mientras que el eje *y* es la frecuencia. Se hallaron los siguientes valores del parámetro σ: Personales (izq), σ = -1,04; Diarios (centro), σ = -0.99 y Empresas (der), σ = -1,07

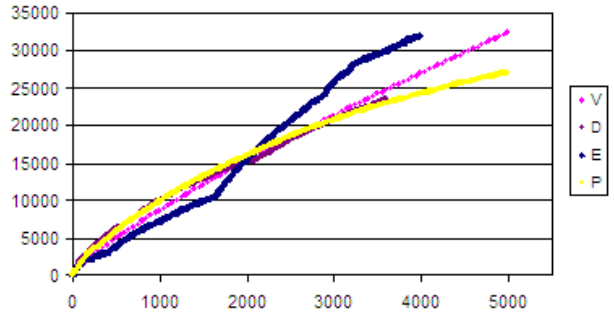

Figura 2 – Distribuciones de crecimiento del vocabulario por categoría con su curva de ajuste a la ley de Heaps. El eje *x* corresponde a la cantidad de *posts*, mientras que el eje *y* es el tamaño del vocabulario.

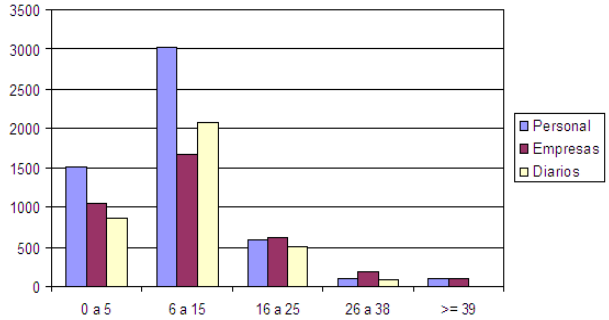

Figura 3 – Distribución de tamaños de *posts* por categoría. En el eje *x* corresponde a la cantidad de términos únicos y en el eje *y,* la cantidad de *posts*.

Por último, se realizó un estudio de la distribución del tamaño de los *posts* publicados – medidos en cantidad de términos – dentro de cada categoría. En la figura 3 se muestran los intervalos analizados. Aquí se puede apreciar que el comportamiento de las tres clases es similar en

cada uno de los intervalos. En los tres casos aproximadamente el 50% de los post publicados contienen entre 6 y 15 términos, esto es coherente ya que se trata de noticias o textos cortos.

4 – Consideraciones Finales

Estos experimentos preliminares, de carácter exploratorio, permiten comparar colecciones formadas con fuentes de información heterogéneas, distribuidas y dinámicas, de acuerdo a los modelos de documentos tradicionales. Las leyes de Zipf y Heaps también son válidas para estas colecciones, aunque el vocabulario crece más rápidamente ($\beta = 0.82$). Estos resultados también se aplican a la definición de las estructuras de datos necesarias en el sistema.

No obstante, los resultados sugieren que algunas de la características deben ser consideradas especialmente a la hora de su incorporación en el algoritmo de selección de recursos. En particular, la longitud de los documentos (*posts*) es un parámetro de importancia debido a sus bajos valores.

5 – Formación de Recursos Humanos

Este trabajo es parte de la tesis de licenciatura del primer autor, la cual 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. Particularmente, este artículo surge como continuación del trabajo presentado en [cita wicc 2007], en el cual se presentó la idea global y los primeros resultados.

6 – Referencias

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

[Banchero et al., 2007] Banchero, Santiago; Bordignon, Fernando R.A. y Tolosa, Gabriel H. 2007. Selección de Recursos Distribuidos en Ambientes Dinámicos Basados en Web. IX Workshop de Investigadores en Ciencias de la Computación. WICC, 2007.

[Callan, 1995] James P. Callan , Zhihong Lu , W. Bruce Croft, 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, July 09-13, 1995.

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

[French et al., 1998] James C. French , Allison L. Powell , Charles L. Viles , Travis Emmitt , Kevin J. Prey, Evaluating database selection techniques: a testbed and experiment, Proc.of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998.

[French et al., 1999] James C. French , Allison L. Powell , Jamie Callan , Charles L. Viles , Travis Emmitt , Kevin J. Prey , Yun Mou, 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, 1999.

[Gelbukh, 2001] Gelbukh, Alexandr, and Sidorov, Grigori. Zipf and Heaps Laws' Coefficients Depend on Language. Proc. CICLing-2001, Conference on Intelligent Text Processing and Computational Linguistics, 2001.

[Hammond et al., 2004] Tony Hammond, Timo Hannay, and Ben Lund. The Role of RSS in Science Publishing. Syndication and Annotation on the Web. D-Lib Magazine. Vol.10 Nº 12, 2004.

[Heaps, 1978] Heaps, H.S. Information Retrieval - Computational aand Theoretical Aspects. Academic Press, 1978.

[O' Reilly, 2007] Tim O' Reilly. 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.

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

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