Modelos y algoritmos de búsqueda + redes sociales para aplicaciones verticales de recuperación de información

Gabriel H. Tolosa, Fernando R.A. Bordignon

Departamento de Ciencias Básicas
Universidad Nacional de Luján
Cruce rutas 5 y 7
{tolosoft, bordi}@unlu.edu.ar

Resumen

El espacio web no es solamente un enorme repositorio de información de todo tipo, sino - además – es una plataforma para soportar servicios globales de naturaleza diversa. El incremento exponencial de contenido y de usuarios (por ejemplo: en las redes sociales), junto con la constante aparición de nuevas aplicaciones, exceden largamente la visión de la web como un mero repositorio de contenidos. En todos los casos, existe como común denominador la necesidad de realizar “búsquedas” de diferente tipo y con objetivos también diversos.

En la actualidad, las redes sociales son unas de las aplicaciones más populares, incluso han modificado la forma en que los usuarios se vinculan, relacionan, interactuan e intercambian información. De forma implícita, generan estructuras sociales con propiedades emergentes que surgen del comportamiento global y, se estima, pueden aportar a mejorar los procesos de búsquedas.

En este documento se presenta un nuevo proyecto de investigación, donde se propone abordar algunas de las problemáticas relacionadas con las búsquedas en Internet. Para ello, se integrarán técnicas de recuperación de información y construcción de motores de búsqueda, junto con información proveniente de redes sociales, para brindar mayor eficiencia en la tarea de búsqueda, abarcando múltiples escenarios como: porciones específicas de la web, información científica y/o geográfica, búsquedas en dispositivos móviles, entre otras.

Palabras clave: motores de búsqueda, redes sociales, búsqueda vertical.

Contexto

Esta presentación corresponde al proyecto de investigación del título, perteneciente al Departamento de Ciencias Básicas de la Universidad Nacional de Luján. El mismo ha sido evaluado externamente, aprobado por Disposición del Consejo Directivo Departamental Nro. 008-11, y cuenta con una vigencia de 2 años.

Este proyecto se vincula – además – con líneas de I+D de sus integrantes en otras instituciones, a saber: Depto. de computación, Universidad de Buenos Aires; LabTIC, Universidad Pedagógica Provincial e Instituto de Clima y Agua, INTA Castelar.

Introducción

En los tiempos actuales, la colaboración y la participación son conceptos que están más vigentes que nunca en la historia de la humanidad. Esto se refleja en los cambios sociales que llevan adelante las tecnologías de la información y las comunicaciones, dado que están influyendo en la construcción del conocimiento. La posibilidad de que “todos” o “casi todos” puedan tener acceso a información de forma instantánea produce una cantidad de efectos en distintas áreas o espacios (democracia, formas de trabajo, relaciones, educación, conocimiento, inteligencia, etc).

En este contexto, el espacio web se ha convertido no solamente en un enorme repositorio de información de todo tipo sino – además – en una plataforma para soportar servicios globales de naturaleza diversa

[Berners-Lee et al., 2000] [Wu, 2002] [Escudeiro et al., 2008]. La última década ha mostrado un enorme crecimiento en la cantidad de información disponible de forma electrónica, no solamente debido a las oportunidades de generación de nuevos datos sino - además - debido al crecimiento y mejora de la eficiencia en las redes de computadoras y algunas aplicaciones "clave" en la red Internet como los motores o máquinas de búsqueda [Cho et al, 2004].

Cada día, millones de usuarios acceden a la red, tanto para consultar fuentes de información como para interactuar con otros, trabajar, estudiar y/o entretenerse. En todos los casos, existe como común denominador la necesidad de realizar “búsquedas” de diferente tipo y con objetivos también diversos [Jaffri et al., 2007]. Sin embargo, en la actualidad ya no resulta suficiente. El incremento exponencial de contenido (de acuerdo a [WWW2010] el tamaño de la web supera los 27 mil millones de páginas) y usuarios (por ejemplo, solamente Facebook [Facebook, 2010] reporta más de 500 millones) junto con la constante aparición de nuevas herramientas y aplicaciones que atraen el interés de los usuarios, y que en algunos casos exceden largamente la visión de la web como un mero repositorio de información textual exigen que se investiguen y desarrollen nuevas ideas, modelo, técnicas y herramientas computacionales, que permitan satisfacer mas eficientemente las necesidades de acceso, tanto desde la perspectiva de tiempo y espacio como de precisión en los resultados [Escudeiro et al., 2008].

En la actualidad, las redes sociales son unas de las aplicaciones más populares, incluso han modificado la forma en que los usuarios se vinculan, interactuan e intercambian información. Por ejemplo, Flickr, Youtube y Facebook [Boyd et al., 2008] [Lewis et al., 2008] atraen a millones de usuarios quienes aportan e intercambian diferentes elementos de contenido (videos, fotos, enlaces, etiquetas). Pero, de forma implícita, generan estructuras sociales con propiedades emergentes [Albert et al., 2002] que surgen del comportamiento global [Xiang et al., 2010]. Otros casos son los sistemas de citas (como Citeulike[1] o Zotero[2]), de comparación de productos y/o subastas (eBay[3]) y los buscadores verticales (Technorati[4]), entre cientos. Además, en un escenario que incorpora semántica el concepto de “buscar algo” quedará redefinido por el de “cumplir objetivos”. El escenario también permite compartir los resultados a partir de su publicación – por ejemplo – en una red social, incrementando el conocimiento global [Motta et al., 2006].

Si se considera que estos servicios son procesos humanos (y no meramente tecnológicos), su mejor comprensión posibilitará aprovechar la inteligencia colectiva para mejorar otros servicios como las búsquedas web y aplicar en nuevos escenarios. La inteligencia colectiva [Lévy, 2004] produce una respuesta simultánea de diversas personas, las cuales están conectadas a un nivel superior de comunicación, trabajando en un mismo proyecto. La integración de toda la información contenida en las redes sociales con las aplicaciones de búsqueda es un tema de alto interés en la comunidad científica relacionada con las áreas de recuperación de información, tecnologías de búsqueda y minería de la web, principalmente [Jaffri et al., 2007].

Finalmente, la búsqueda se convirtió en un proceso “central”. Los datos son cada vez mas “ricos”, complejos y ocurren en tiempo real, aportando nuevo valor, pero solamente si están disponibles en tiempo y forma [Hall et al., 2009]. Esta problemática tiene aún muchas preguntas abiertas y – mientras se intentan resolver cuestiones – aparecen nuevos desafíos, lo que requiere permanentemente mejorar la capacidad de los sistemas para realizar esas tareas en forma eficiente (como se mencionó, tanto en performance como en calidad de los resultados) o diseñar nuevas aplicaciones.

Líneas de investigación y desarrollo

La investigación propuesta tiene un balance teórico y experimental, conviviendo el

1 http://www.citeulike.com/
2 http://www.zotero.com/
3 http://www.ebay.com/
4 http://www.technorati.com/

modelado y análisis con el desarrollo y pruebas empíricas de algoritmos y métodos. Por otro lado, hay una componente importante de desarrollo de aplicaciones concretas (al menos como prototipos) que presenten las soluciones a los problemas planteados. En particular, se orienta el proyecto a dos áreas principales, integrando ambos mundos:

a) **Motores de búsqueda**

Como se ha mencionado, los motores de búsqueda son aplicaciones fundamentales en Internet. Su diseño plantea desafíos de naturaleza diversa, desde la recolección e indexación eficiente del contenido hasta cuestiones de escalabilidad. En este sentido, se pretende trabajar en dos sub-areas:

- La arquitectura para aplicaciones específicas, diseñando aplicaciones de búsqueda ad-hoc para problemas concretos, donde una solución de propósito general no es la más eficiente. Aquí se deben estudiar cómo los diferentes modelos de distribución de documentos e indexación determinan la eficiencia del sistema

- El *front-end* de los motores de búsqueda existentes. Hoy en día, los proveedores de servicios de búsqueda brindan interfaces de aplicación (APIs) para la utilización de sus bases de datos. Se propone utilizar los resultados de aplicaciones de búsqueda clásicas para retroalimentar modelos de post-procesamiento que permitan mejorar la respuesta, por ejemplo, utilizando información de perfiles de usuarios, la caracterización de determinados sitios o la historia de navegación anterior, entre otras.

b) **Redes sociales**

Comprender las propiedades de las estructuras que se forman y el comportamiento de los usuarios a los efectos de proveer de evidencias que permitan mejorar las búsquedas. Aquí, se plantea la definición de concepto de "*Query experto*", a partir de comprender cómo usuarios de redes sociales interactúan y valoran los procesos de búsqueda. La intención es integrar este concepto con aplicaciones de *front-end* de los motores de búsqueda anteriormente expuesta,

Por otro lado, la comprensión de cómo se forman las redes sociales y cómo evolucionan [Burger et al., 2009] [Kossinets et al., 2006] posibilita diseñar algoritmos y topologías eficientes para compartir y distribuir la información generada. Esto es especialmente interesante si se tiene en cuenta que la red es un ambiente altamente dinámico y de gran escala.

En ambos casos, se enfoca la investigación la aplicación de algoritmos de búsqueda sobre tipos particulares de contenidos, lo que se conoce como búsqueda vertical. El concepto de "tipo" puede hacer referencia tanto a una especialización temática (educación, agricultura), funcional (ventas, subastas) o clase de contenido (fotos, videos, música). La especialización permite reducir – además – la problemática de la escalabilidad. De otra forma, se requeriría de una gran poder de cómputo si se pretende operar a escala de la web.

Resultados y Objetivos

El objetivo principal es estudiar, desarrollar, aplicar, validar y transferir modelos, algoritmos y técnicas que permitan construir herramientas y/o arquitecturas para abordar algunas de las problemáticas relacionadas con las búsquedas en Internet [Henzinger, 2005]. Los aportes directos de este proyecto están relacionados con la integración de técnicas de recuperación de información y construcción de motores de búsqueda junto con información proveniente de redes sociales digitales para construir aplicaciones específicas que brinden mayor eficiencia a los usuarios en la tarea de búsqueda. Dichas aplicaciones pueden abarcar múltiples escenarios (verticales) como: porciones específicas de la web, información científica, información geográfica, búsquedas en dispositivos móviles, entre otras.

Específicamente, se espera:

- Generar mejoras en las estrategias y posibilidades que brindan los motores de búsqueda utilizando información de contexto para construir aplicaciones específicas de recuperación de información.
- Diseñar y construir interfaces de búsqueda avanzadas como *front-end* de buscadores clásicos
- Diseñar y construir aplicaciones de búsqueda basados en un enfoque vertical (por ejemplo, en dispositivos móviles) que permitan aprovechar las características particulares de la información en cuestión.
- Caracterizar y establecer las relaciones y asociaciones existentes entre los elementos de información de los dominios mencionados.

Complementariamente, la temática puede aportar soluciones concretas para resolver problemas del mundo real con lo que se estiman múltiples oportunidades de transferencia a la sociedad.

Formación de Recursos Humanos

En el marco de la temática de este proyecto, y como continuación de trabajos previos, se están dirigiendo dos trabajos finales de grado de la Licenciatura en Sistemas de Información de la UNLu. Uno de los integrantes del proyecto está comenzando en la docencia (auxiliar alumno) e investigación en la Universidad y – además – se cuenta con un pasante (estudiante avanzado) que comenzó a participar como asistente en investigación.

Por otra parte, uno de los integrantes del proyecto está realizando el Doctorado en Ciencias de la Computación de la Universidad de Buenos Aires y otro ha comenzado sus estudios de postgrado en el área de minería de datos en la misma institución. Un tercer integrante se encuentra finalizando la Maestría en Tecnologías Integradas y Sociedad del Conocimiento en la Universidad Nacional de Educación a Distancia de España.

Referencias

[Albert et al., 2002] R. Albert and A-L. Barabasi. Statistical mechanics of social networks. Reviews of Modern Physics, 74. 2002.

[Baeza, 2003] R. Baeza-Yates. Information Retrieval in the Web: beyond current search engines, International Journal on Approximated Reasoning 34 (2-3), pp: 97--104, 2003.

[Baeza-Yates et al, 2005] R. Baeza-Yates, C. Castillo, V. López. Characteristics of the web of Spain. Cybermetrics - International Journal of Scientometrics, Informetrics and Bibliometrics , 9. 2005

[Baeza-Yates et al, 2007a] R. Baeza-Yates, C. Castillo, E. Efthimiadis. Characterization of national web domains. ACM Transactions on Internet Technology, 7. 2007.

[Baeza-Yates et al, 2007b] R. Baeza-Yates, C. Castillo, E. Graells. Características de la web chilena 2006. Technical report, Center for Web Research, University of Chile. 2007.

[Berners-Lee et al., 2000] T. Berners-lee, M. Fischetti, M.L. Dertouzos. T. Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web. HarperCollins Publishers. 2000.

[Boyd et al., 2008] D. Boyd, N. Ellison. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1):210–230, 2008.

[Burger et al., 2009] M. Burger and V. Buskens. Social context and network formation: An experimental study. Social Networks, 31(1):63–75, January 2009.

[Cho et al, 2004] Cho, J. & Roy, S. Impact of search engines on page popularity. In Proceedings of the 13th international conference on World Wide Web, ACM Press, 2004.

[Escudeiro et al., 2008] N.F. Escudeiro, A. M Jorge. Satisfying Information Needs on the Web: a Survey of Web Information Retrieval. Polytechnical Studies Review , Vol VI, no 9 . 2008.

[Facebook, 2010] Facebook Datos Estadísticos (consulta: 30/07/2010). http://www.facebook.com/press/info.php?statistics

[Gomes et al., 2005] D. Gomes, M.J. Silva. Characterizing a national community web. ACM Transactions on Internet Technology, n. 5, 2005.

[Hall et al., 2009] W. Hall, D. De Roure, N. Shadbolt. The evolution of the Web and implications for eResearch. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, Vol. 367, No. 1890, pp. 991-1001. 2009.

[Henzinger, 2005] M. Henzinger, M. Algorithmic challenges in web search engines. Journal of Internet Mathematics, n.1, 2005.

[Jaffri et al., 2007] A. Jaffri, H. Glaser. Knowledge Enhanced Searching on the Web. In 6th International Semantic Web Conference, Doctoral Consortium, pp: 921—925, Busan, Korea. 2007.

[Kossinets et al., 2006] G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. Science, 311(5757):88–90, January 2006.

[Kumar, 2006] Kumar, R. New search paradigms (session). Proceedings of the 15th international conference on World Wide Web, Edinburgh, Scotland, 2006.

[Kumar et al., 2009] R. Kumar, A. Tomkins. A Characterization of Online Search Behavior. IEEE Data Engineering Bulletin, Vol.32, n2, pp: 3--11, 2009

[Levene et al., 2003] M. Levene, A. Poulovassilis (editors). Web Dynamics, Springer, 2003.

[Lévy, 2004] P. Lévy.. Inteligencia Colectiva: Por una antropología del ciberespacio. Traducción por: Felino Martínez Álvarez, Organización Panamericana de la Salud, Washington DC, pp. 14-148, 2004

[Lewis et al., 2008] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer and N. Christakis. Tastes, ties, and time: A new social network dataset using facebook.com. Social Networks, 2008.

[Motta et al., 2006] E. Motta, M. Sabou. Next Generation Semantic Web Applications. In Proc. of the 1st Asian Semantic Web Conference (ASWC}, pp 24--29, Springer, 2006

[Sanderson et al., 2007] M. Sanderson, S. Dumais. Examining repetition in user search behavior. Proceedings of ECIR ’07, 2007.

[Tolosa et al., 2007] G. Tolosa, F. Bordignon, R. Baeza-Yates, C. Castillo. Characterization of the Argentinian Web. Cybermetrics, Vol. 11 , Issue 1, Paper 3. 2007.

[Xiang et al., 2010] R. Xiang, J. Neville and M. Rogati. Modeling Relationship Strength in Online Social Network. 19th International World Wide Web Conference, 2010.

[Wu, 2002] W. Hu. World Wide Web Search Technologies, Architectural Issues of Web-Enables Electronic Business, Idea Group Publishing . 2002.

[WWW2010] World Wide Web Size.com (consulta: 30/07/2010). http://www.worldwidewebsize.com/