B.3. Nueva generación de motores de búsqueda basados en procesamiento de lenguaje natural

Por **José-Ramón Pérez-Agüera**

Pérez-Agüera, José-Ramón. "Nueva generación de motores de búsqueda basados en procesamiento de lenguaje natural". En: *Anuario ThinkEPI*, 2008, pp. 39-40

Resumen: *Breve análisis de la situación actual del mercado de buscadores basados en procesamiento de lenguaje natural. Powerset y Hakia son puestos en la palestra como máximos exponentes de esta tendencia.*

Palabras clave: *Buscadores*, *Lenguaje natural*, *Powerset*, *Hakia*

Title: The new generation of search engines based on natural language processing

Abstract: *Current trends in natural language processing and information retrieval are briefly analysed to show their impace on the search engine market. Powerset and Hakia are highlighted as the main agents of this change.*

Keywords: *Search engines*, *Natural language*, *Powerset*, *Hakia*

EN EL ÚLTIMO AÑO se ha visto florecer toda una serie de nuevos buscadores cuya característica común ha sido la integración de técnicas de procesamiento de lenguaje natural en el proceso de búsqueda.

Los dos baluartes de esta nueva tendencia son *Powerset*[1] y *Hakia*[2], detrás de los cuales se ha reunido la *crème de la crème* del procesamiento de lenguaje natural para conseguir un nuevo salto de calidad en la evolución de los buscadores web.

La idea de integrar conocimiento lingüístico en los buscadores no es nueva en absoluto, y desde los años 90, si no antes, se han repetido los intentos de implementar buscadores que fueran más allá de recuentos más o menos complicados de frecuencias de palabras. El más sonado fracaso a este respecto fue sin duda el intento de **Ellen Voorhees**, allá por 1993, de usar *Wordnet*, una base de datos enorme con información semántica para expandir las consultas de los usuarios.

Los resultados de este experimento, como se puede ver en su artículo[3] fueron bastante desoladores y desde entonces, más allá de estudios puntuales cuyos resultados no han sido concluyentes, el uso de lenguaje natural en recuperación de información ha quedado relegado a la aplicación de técnicas bastante triviales como el *stemming* y la eliminación de palabras vacías.

La razón de este nuevo resurgimiento del lenguaje natural en el entorno de los buscadores se corresponde en parte con un ciclo natural, típico de cualquier disciplina científica, donde se prueban viejas ideas desde enfoques nuevos. Pero también se trata de una cuestión de mercadotecnia, donde nuevos buscadores tratan de entrar en el mercado vendiendo la idea de que tienen una nueva tecnología revolucionaria que superará con creces el enfoque actual de los grandes buscadores.

Desde el punto de vista científico, el león no es tan fiero como lo pintan, y al igual que

La razón del resurgimiento del lenguaje natural en el entorno de los buscadores responde en parte a un ciclo natural pero también se trata de una cuestión de mercadotecnia

Powerset y *Hakia* han puesto a trabajar a importantes investigadores en procesamiento de lenguaje natural, *Google*, *Yahoo* y *Microsoft* llevan tiempo trabajando también en esa dirección.

La conclusión que se puede sacar de aquí es que, pese a que la inclusión de lenguaje natural en los buscadores es sin duda una de las líneas de trabajo futuro para mejorar no sólo la calidad de los resultados de los buscadores sino también sus posibilidades e interacción con los usuarios, aún queda mucho por hacer a este respecto, y raro será que ningún nuevo buscador desbanque a *Google* simplemente por utilizar técnicas de procesamiento natural.

En este sentido, se ha de ser conscientes que el *boom* de *Google* en 1998 estuvo más relacionado con su entrada en un mercado prácticamente virgen respaldado por una fuerte inversión económica que con una ventaja tecnológica decisiva, ya que sin menospreciar la importancia del *Pagerank* es importante recordar que no eran los únicos que usaban un algoritmo de análisis de enlaces.

Pese a todo lo dicho, merece la pena seguir los avances que se hagan a este respecto: tanto aquellos que vengan de ultramar, como lo que se desarrollen en Españа por ejemplo de la mano de empresas como *Bitext*, no vaya a ser que un día nos sorprendamos de las maravillas lingüísticas que son capaces de hacer los buscadores americanos sin saber que aquí

Raro será que ningún nuevo buscador desbanque a Google simplemente por utilizar técnicas de procesamiento natural

cerca tenemos una empresa española que es la que hace posible esas maravillas.

Notas

1. *http://www.hakia.com*

2. *http://www.powerset.com*

3. **Voorhees, E. M.** "Using WordNet to disambiguate word senses for text retrieval". En: *Proceedings of the 16th Annual international ACM Sigir Conference on Research and Development in information Retrieval,* 1993, pp. 171-180.
http://doi.acm.org/10.1145/160688.160715