B.10. Recuperación de información en las empresas

Por **José-Ramón Pérez-Agüera**

Pérez-Agüera, José-Ramón. "Recuperación de información en las empresas". En: *Anuario ThinkEPI*, 2008, pp. 67-70.

Resumen: *Crónica del workshop realizado por la empresa Atos Origin en Madrid sobre el mercado de los buscadores. Se contó con la presencia de los principales actores a nivel español, desde Google hasta la Universidad Nacional de Educación a Distancia.*

Palabras clave: *Buscadores*, *Empresa*, *Workshop.*

Title: Information recovery in business and industry

Abstract: A *workshop on the search engine market, organized by Atos Origin in Madrid, is described. The key players in Spain's market, from Google to the national distance learning university, attended the meeting to show their products and to comment on their view of the marketplace for web and enterprise search engines.*

Keywords: *Search engines*, *Enterprise search*, *Workshop.*

EL MIÉRCOLES 26 DE SEPTIEMBRE tuve la oportunidad de asistir a un *workshop* organizado por el departamento de *Media Events* de *Atos Origin* sobre buscadores. El evento fue lo suficientemente interesante como para merecer una nota *ThinkEPI*, no sólo por lo que dijeron sino también por lo mucho que no se dijo y por las reflexiones que de ello pueden derivar.

La mañana se dividió en dos partes: la primera moderada por **Isidro Aguillo**, que se encargó magistralmente de que nuestra profesión estuviera bien representada; y la segunda moderada por **Julio Gonzalo**, profesor titular de la *Uned* y uno de los investigadores españoles más importantes en *Recuperación de Información* y *Procesamiento de Lenguaje Natural*.

En el primer bloque de charlas, todas ellas de unos 25 minutos de duración, pudimos escuchar un entretenido monólogo a **Adolfo Corujo**, responsable del departamento de marketing del grupo *Alma*, quien, al más puro estilo de "El Club de la Comedia", amenizó la mañana con anécdotas y chascarrillos procedentes de los distintos proyectos de integración de su solución de búsqueda en entornos empresariales. La idea de partida era señalar todas aquellas cosas que les piden los clientes cuando integran un buscador en su sistema y que un buscador no puede hacer.

De esta charla surgió la primera de las reflexiones que he tenido a partir de ese día: "Muchos de los responsables de sistemas de información de las empresas no tienen ni idea de lo que es un buscador ni de para qué sirve", de donde se deduce que tampoco tienen claro qué problemas pueden y no pueden solucionar con él. Y la verdad es que esto me parece grave: ya que el hecho de que mi madre no sepa lo que es un buscador, no es un problema; pero que la persona encargada de elegir un producto, y por lo tanto de gastarse una gran cantidad de dinero, ya sea en una institución pública o privada, no tenga claro qué es y para qué sirve una herramien-

El uso del procesamiento del lenguaje natural y las ontologías para recuperación de información está sobrevalorado, y que sólo sirve para dominios concretos

ta de búsqueda me parece absolutamente catastrófico.

Quizás suene extraño esto que digo, pero pongamos un ejemplo más prosaico.

Imaginemos que entramos en una tienda a comprar un televisor, uno de esos planos que tanto gustan y que últimamente se está comprando todo el mundo. Después de mirar un rato por los expositores damos con una tele de aspecto evocador, moderna y que hará que nuestro salón parezca la oficina de Tom Cruise en *Minority Report*.

Convencidos de que eso es lo que necesitamos, no sólo porque nos hará más felices, sino porque nuestro vecino ya tiene una y nosotros no vamos a ser menos, nos acercamos al dependiente y le preguntamos por sus características. El abnegado trabajador nos responde con una serie de especificaciones técnicas que no entiende nadie. Pero no importa, necesitamos ese aparato; así que le preguntamos por un par de detalles para asegurarnos de que nuestra felicidad será completa:

–¿Se puede ver la *TDT*?

– Por supuesto, lleva un decodificador incorporado –nos responde el dependiente.

"Genial", pensamos mientras nos deleitamos con la calidad de la imagen del *DVD* de alta definición que está reproduciendo nuestra futura tele en el expositor. –Caramba, y se ve muy bien, oiga –le decimos al dependiente.

– Claro –nos responde– está adaptada para los nuevos sistemas de alta definición.

No sabemos qué es eso, pero qué bien suena. Tras unos breves momentos de pensar lo que dirá la pareja por habernos dejado 1.800 euros en un televisor de 42' que ya veremos cómo lo ponemos en nuestro minúsculo pero altamente tecnificado salón, sacamos la *Visa* y que arda Roma, que la vida son dos días.

Una vez en nuestro hogar conectamos nuestra primorosa televisión y la encendemos esperando ver aquellos deliciosos paisajes del *DVD* de la tienda con una calidad que hace que te sientas como si estuvieras allí. Entonces se da la primera decepción: "Vaya, esto se ve fatal, a ver si va a estar rota". Cogemos la factura y nos bajamos a la tienda.

Al entrar el dependiente nos mira con cara de circunstancias, intuyendo que va a tener el primer problema del día.

La Web Semántica ha traído y traerá muchas cosas buenas a la Red, aunque no mejorará la recuperación de información en internet por problemas estructurales básicos

– Oiga mi tele se ve fatal –le decimos intentando mantener la calma

– ¿La ha sintonizado correctamente?

– Pues sí –le decimos con ganas de añadir que no somos tontos y que sabemos sintonizar un televisor con la ayuda del manual.

Entonces el señor nos explica que la emisión, incluso de la *TDT*, no es de alta definición y que por lo tanto no podemos esperar que se vea igual que el *DVD* que nos pusieron en la tienda.

– Oiga, pero es que se ve peor que la tele que tenía antes.

A lo que nos responde que eso se debe a que los televisores analógicos se ven mejor con señales de peor calidad debido a que nuestra televisión al tener más resolución no puede llenar todos los píxeles de la pantalla con una señal de tan baja calidad.

Conclusión: nuestra tele se ve peor porque es tan sofisticada que si la señal no es igual de buena no funciona correctamente. Explícale eso a tu pareja, que te va a estar sacando el tema del televisor hasta Navidades (y eso que todavía no ha visto la factura de la *Visa*).

Si transportamos este ejemplo al caso de los buscadores tendremos un problema parecido, sustituyamos la tele por un buscador y a la pareja por nuestro jefe y ya la tenemos montada.

Conclusión: cuando compramos de oídas, pensando que sabemos lo que compramos y sin tener claro para qué lo vamos a usar, lo normal es que nos surjan infinidad de problemas aunque compremos un producto de gran calidad.

Solución: formación, formación y formación, y menos caras raras cuando el profesor nos explique el modelo de espacio vectorial o *BM25* en clase pensando que eso son cosas de informáticos.

En fin, como no se puede tener todo, pasemos a las charlas siguientes, prometo no extenderme tanto.

El siguiente en hablar fue **Monte Kluemper**, de *BEA Systems*, que nos dio una charla completamente olvidable sobre las soluciones de búsqueda que ofrece su empresa (prescindibles).

La última charla de este primer bloque la brindó **Aljosa Pasic**, de *Atos Research & Innovation*. Dado que *Atos* se encargó del desayuno del *workshop* no criticaré su charla, no sea que no me vuelvan a invitar, y me limitaré a decir que la presentación fue más bien floja.

Después del café, comenzó el segundo bloque con la intervención de **Hugo Zaragoza**, de *Yahoo! Research*, sobre la web 2.0 y la fuerza de la comunidad. La charla estuvo bien, pero yo no podía dejar de pensar que llamar a uno de los mayores expertos que hay en el mundo en funciones de *ranking* probabilísticos para que hable de los *blogs*, la *Wikipedia* y *Yahoo respuestas* es un desperdicio.

El siguiente en hablar fue **Stefano Aldrovandi**, de *Exalead*, quien nos deleitó con una entretenida presentación sobre las funcionalidades y usabilidad de su buscador: nada del otro mundo en cuanto a contenido, pero muy bien presentado.

Francisco Serrano-Maestre, de *Microsoft,* fue el encargado de la siguiente charla. Comentaría qué tal estuvo, pero fue demasiado densa.

Tan sólo resaltar una cosa que me llamó la atención, y es la obsesión de las empresas en general por vender chapuzas. Todo lo resuelven; para ellos no hay problemas, sino soluciones. Luego cuando se escarba un poco se ve que todas sus soluciones están cableadas, es decir, resueltas de mala manera para que funcionen en unos pocos casos concretos si, y sólo si, se cumplen unas premisas muy estrictas. Por ejemplo, una de las fantásticas funcionalidades de su solución de búsqueda es que era capaz de extraer el organigrama de una empresa y permitirte buscar nombres de personas en tu colección de documentos. "Vaya, reconocimiento automático de entidades nombradas", podría pensar uno. Pues no, sólo eran capaces de hacerlo a partir de un directorio activo, que no es otra cosa que una base de datos de personas. Yo también soy capaz de buscar nombres de personas cuando tengo un diccionario con los nombres que quiero buscar. Lo complicado es extraer

nombres de personas de los documentos sin que aparezcan identificados explícitamente.

Esto me hizo pensar en el hecho de que, en muchas ocasiones, las empresas venden humo diciendo que resuelven problemas que realmente no son capaces de resolver. Este problema también se solucionaría si el responsable de comprar estos productos supiera de lo que habla y fuera capaz de hacer las preguntas correctas antes de comprar una solución de búsqueda.

La siguiente presentación corrió a cargo de **Carlos García Armendáriz**, responsable de *enterprise search* de *Google*, o lo que es lo mismo, el que vende *Google Search Appliance* a las empresas.

El problema que tiene *Google* en España, y se vio muy claramente en esta charla, es que como no tienen competencia pasan de hacer ningún tipo de esfuerzo por vender. Esto da como resultado, que aquél puso una diapositiva en blanco con el logo de *Google* en el centro y casi ya está todo dicho.

Nos explica lo que luce su buscador, lo estupendos que son, y que no tienen problemas de escalabilidad porque han indizado toda la Web y se queda tan tranquilo. Era para haberle dicho que su *Pagerank* en una intranet no sirve de mucho, y que su solución de búsqueda no está por encima del resto de las que hay en el mercado debido a que a día de hoy todas implementan la misma tecnología.

La penúltima charla corrió a cargo del responsable de sistemas de la *Agencia EFE*. Fue como ver un capítulo de "Cuéntame como buscó": nos habló de toda la historia de automatización de su archivo desde los años 80 hasta la actualidad. Nada reseñable más allá de que en mi opinión podrían usar *software*

Julio Gonzalo, Universidad Nacional de Educación a Distancia.

libre y les saldría todo mucho más barato.

El último en hablar fue **Julio Gonzalo**, de la *Uned*. Estuvo realmente genial, pues habló de semántica y de Web Semántica, y dijo todas aquellas cosas que yo llevo deseando oír desde hace mucho tiempo a un profesional de su nivel:

– Que el uso del procesamiento del lenguaje natural y las ontologías para recuperación de información está sobrevalorado, y que sólo sirve para dominios concretos.

– Que la Web Semántica le ha traído y le traerá muchas cosas buenas a la Red, como por ejemplo la sindicación de contenidos; y que, sin embargo, no mejorará la recuperación de información en internet por problemas estructurales básicos, como el hecho de que no se puede pretender que los usuarios etiqueten sus páginas web cuando muchos ni siquiera les ponen el título.

La verdad es que **Julio** estuvo muy bien y fue una pena que no tuviera tiempo par dar su charla completa.

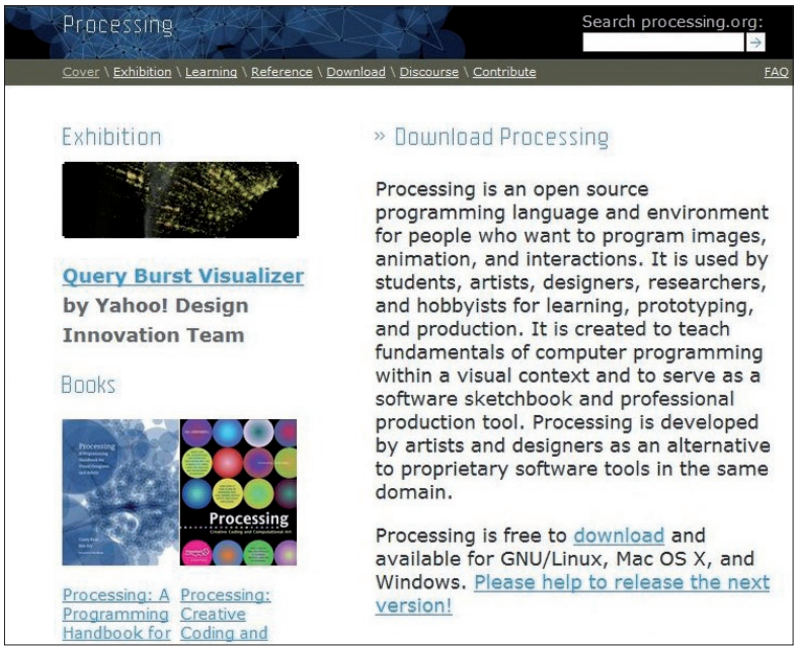
En definitiva, una jornada interesante sobre todo para los que tenemos un perfil más académico, como es mi caso, y vivimos en nuestra torre de marfil sin pensar mucho que las cosas sobre las que investigamos luego son usadas por gente tanto en sus casa como en su trabajo.

Buscadores en intranets

Por **Óskar Calvo-Vidal**

Cierto es que a todas estas charlas, si es cierto que la mayoría de ellas son infumables, habría añadido una más para hablar de las soluciones que puede aportar *Lucene/Nutch* a la búsqueda de contenidos en intranets. Es verdad que son soluciones de *software* libre, pero esto no quita para que sean formas de negocio (recordemos que las implantaciones de s.l. se siguen facturando a los clientes).

Por otro lado, y unido a *Lucene/Nutch*, el hecho de poder usar una herramienta como *Processing* para jugar con la información es una solución interesante, sería una forma de presentar en *software* libre lo mismo que están haciendo los chicos de *Aquabrowser*.



http://processing.org/

Volviendo al tema de los jefes de departamentos informáticos, suelen ser muy monolíticos; muchos de ellos opinan que por tener una empresa que tiene el *"partner"* de *Microsoft* están contratando lo mejor, y no les entra en la cabeza otras soluciones.

Muchas veces suele depender también de estas personas el desarrollo de las herramientas de los departamentos de documentación: su desconocimiento en la materia significa la compra de "engendros" que normalmente no sirven para nada, pero que por el precio que se ha pagado hay que apechugar con ellos.