# Reseñas

## Jotri'2002: Jornadas de tratamiento y recuperación de información

Por **José Antonio Ontalba** y **Susana Serrano**

**LA *UNIVERSIDAD POLITÉCNICA DE VALENCIA* organizó las primeras *Jornadas de tratamiento y recuperación de información* (*Jotri'2002*), celebradas durante el 4 y 5 de julio de 2002 en Valencia.**

Siguiendo la línea de otros eventos desarrollados en España últimamente, tales como las *Jornadas de bibliotecas digitales* (*Jbidi*), *Jotri'2002* se convocó con el ánimo de aunar esfuerzos y líneas de actuación entre la biblioteconomía y la documentación, de un lado, y el área de procesamiento del lenguaje natural (que engloba la lingüística y la ingeniería informática), de otro.

Con un total de asistentes entre los cien y los ciento cincuenta, se buscaba plasmar el estado actual de los desarrollos relativos al tratamiento y recuperación de la información, así como abrir vías de colaboración en investigación teórica y aplicada.


### Contenidos

Las jornadas no tuvieron ningún subtítulo que indicara el enfoque del contenido que se iba a tratar. Al contrario, su estructura estaba compuesta por cinco sesiones temáticas y dos demostraciones comerciales, además de presentaciones de proyectos que, en resumen, se pueden englobar en dos grandes líneas:

**1. Tratamiento de la información**

Uno de los temas abordados fue **xml**, cuya característica más importante es su capacidad de trabajar sobre las estructuras y los contenidos de los documentos en el entorno web, a diferencia de html, que incide únicamente sobre el formato del documento. Esta particularidad permite recuperar de manera más relevante y precisa la información.

En este ámbito se dieron a conocer trabajos sobre sistemas que comparan la pregunta con la estructura definida en el documento: la técnica aplicada consiste en considerar una pregunta expresada en xml y traducirla a un documento denominado "plantilla", que representa la estructura y el contenido y a través de la cual se obliga a que un documento satisfaga la consulta con una respuesta.

**«A pesar del gran avance de xml, el gran reto es superar los problemas de desambiguación existentes en el ámbito de la recuperación de la información»**

A pesar de los avances en xml, en *Jotri'2002* se reconoció que aún están por superar los problemas de desambiguación existentes relacionados con la recuperación de información.

En lo referente a sistemas automatizados de **resúmenes** se constataron las siguientes mejoras:

—Aplicación de estos sistemas a otros idiomas, especialmente inglés y catalán.

—A través del uso de estándares como *EuroWordNet* se ha alcanzado una mejora en el procedimiento léxico de cara a un mejor

## La visión del informático

***¿Los avances en recuperación de información se pueden beneficiar si trabajan conjuntamente informáticos y documentalistas?***

—Sin duda, los primeros viendo el aspecto técnico y los segundos viendo el aspecto humano, que incluye semántica, preocupación por la persona, etc.

Aunque hay desarrollos multidisciplinarios, faltan más estudios experimentales con usuarios inexpertos que lamentablemente son difíciles y caros de hacer.

***¿Cree que, en el futuro, habrá que pagar para obtener buenos resultados en la búsqueda de información en la web?***

—Es muy posible que sí, sobre todo en búsqueda de información especializada.

Es difícil también mantener un modelo totalmente gratuito al que parece estar acostumbrada la gente (aunque a mí me gustaría también que siguiera así).

Tal vez será algo híbrido, una parte genérica gratis, y otra mejor (o con el contenido mismo) de pago. Es decir, poder encontrar algo pero no verlo completo, sólo un resumen, al estilo de muchas publicaciones electrónicas.

La otra alternativa es que se pague por aparecer en el resultado y por ende la búsqueda siga siendo gratis (es decir, como las paginas amarillas).

***¿Cuál cree que será la tendencia del futuro: simplificar la manera de hacer preguntas (lenguaje natural) o permitir búsquedas avanzadas cada vez más desarrolladas?***

—Ambas, y posiblemente otras como buscar en base a un ejemplo (documentos similares). No creo que sea posible eliminar la ambigüedad del lenguaje natural, pero para usuarios no especializados será la tendencia. Para nosotros, tanto informáticos como documentalistas, siempre preferiremos mejores herramientas.

Ricardo Baeza-Yates

Por otro lado, cuando aumente el uso de xml podremos tener lenguajes que permitan consultar contenido en el contexto de la estructura y por ende mejorar la relevancia de los resultados. *XQuery* será una de estas alternativas, la cual deberá tener interfaces más amigables para el usuario final (por ejemplo lenguajes visuales de consulta donde se dibuja la estructura que se está buscando).

***¿Podemos esperar que pronto haya un* Google *para la internet invisible?***

—Para mi internet invisible es la que está detrás de cortafuegos o de páginas de acceso con contraseña; así, por ejemplo, un fichero no indizado que tiene un formato extraño no lo es porque no sea visible, sino porque no es procesable.

Partiendo de esas premisas, no creo que se pueda esperar tal *Google*: no porque no pueda ser parcialmente posible, sino porque puede no tener mucho sentido. También depende de qué sea la internet invisible, pues parte de esa internet es la web privada y esa seguirá siendo privada a menos de que haya un esquema de sitios web buscables pero no visibles a menos que uno tenga acceso (en algunos casos pagando). La otra web invisible es la web dinámica, que incluye todas las bases de datos accesibles por medio consultas y/o *clicks*. Los *clicks* se pueden seguir, las consultas tal vez se pueden hacer (lo que se llama extracción automática de datos), pero puede ser que se indexen miles de millones de paginas dinámicas de consultas y/o *clicks* que nadie nunca hará y que por ende, al no tener interés, tampoco nadie encontrara.

***¿Cree que la web semántica es una realidad a corto plazo o todavía tendremos que esperar muchos años?***

—Especialmente los temas de deducción lógica y confianza parecen muy lejanos aún.

Lo que creo que será una realidad en el corto plazo es un buen nivel de interoperabilidad semántica, esto es, tener una web donde sea posible buscar, clasificar y comprender significados. Esta es la opinión de uno de mis colegas, **Claudio Gutiérrez**, que trabaja en este tema. Igual puede que esté siendo optimista.

Yo tiendo a ser más negativo, pues hasta ahora el uso de metadatos se ha desvirtuado para confundir a los buscadores y hacerles creer que hay páginas más importantes que otras o que contienen información que en realidad no contienen.

***¿Puede indicar tres tendencias de futuro en la búsqueda de información en la web para los próximos cinco años?***

—Extracción de información, es decir, minería de la web, encontrar cosas que son interesantes pero que no necesariamente estoy buscando a priori.

Mejor búsqueda en documentos multimedia, en particular imágenes y música.

Uso de tecnologías móviles y sin hilos con agentes activos que buscan información personalizada y alertan cuando encuentran algo.

## La visión del documentalista

***¿Los avances en recuperación de información se pueden beneficiar si trabajan conjuntamente informáticos y documentalistas?***

—No solamente se pueden beneficiar, sino que es im-pres-cin-di-ble que trabajemos juntos. De hecho, los problemas de la recuperación son tan complejos que resulta de todo punto necesaria la colaboración de profesionales, de puntos de vista, de intereses y de tradiciones culturales de los dos colectivos. Los documentalistas, con nuestra visión de "alto nivel", tendemos a ignorar los enormes problemas de ingeniería; pero los informáticos con su visión enfocada a la resolución de cuestiones prácticas (la visión de "bajo nivel" en el propio argot informático) tienden a ignorar los enormes problemas cognitivos, lingüísticos y culturales del problema general de "encontrar información sobre algo", cuando ese algo va más allá del estado de una cuenta bancario o el estado de un proceso de compra/venta.

***¿Cree que, en el futuro, habrá que pagar para obtener buenos resultados en la búsqueda de información en la web?***

—Ya sucede, en parte. Si alguien quiere obtener muy buenos resultados a través de la web ya existen servicios que proporcionan esa información solamente a miembros de determinados colectivos. Las mejores bases de datos de publicaciones académicas y profesionales, que se pueden acceder ya a través de internet, siguen siendo de pago.

Al mismo tiempo, creo que el fenómeno de la gratuidad en internet será algo permanente. El cambio real no consistirá en que todo será de pago algún día, sino que habrá una convivencia de los dos modelos: sistemas de pago convivirán con sistemas gratuitos, es decir, en este último caso, sistemas financiados de forma distinta al pago por suscripción o al *pay per view*. Si nos fijamos bien, no es ninguna novedad: es el modelo de la comunicación audiovisual. Las cadenas de pago conviven con las cadenas gratuitas, es decir, que se financian con publicidad; y no es necesario recordar el tema de las publicaciones impresas gratuitas que hace tiempo que están viviendo un auténtico *boom*.

Lluís Codina

Por tanto, imagino un panorama donde servicios como *Bubl*, *Sosig* o *Cite Seer* (gratuitos y de enorme calidad) convivirán con servicios como *Science Direct*, *Ieee Explore* o *Wiley Interscience* (de pago y de enorme calidad), pero cada uno con una oferta complementaria. Es por eso que algunos hablamos de ecosistema de la información en la web.

***¿Cuál cree que será la tendencia del futuro: simplificar la manera de hacer preguntas (lenguaje natural) o permitir búsquedas avanzadas cada vez más desarrolladas?***

—Lo que busca todo el mundo es simplificar y permitir la manera de hacer preguntas en lenguaje natural, pero como eso no resulta posible, al menos de forma eficiente en todos los casos, todo el mundo acaba desarrollando interfaces de búsqueda avanzada.

Un ejemplo paradigmático aquí es *Google*. Cuando empezaron, los de *Google* confiaban tanto en su sistema de relevancia que no tenían formulario de búsqueda avanzada. Creían que bastaba con usar el operador *AND* de forma implícita y no había una opción de búsqueda avanzada. Con el tiempo, vieron una cosa que los documentalistas ya sabíamos desde hace tiempo, por nuestra estrecha relación con los usuarios finales: no todas las necesidades de información se pueden representar con un *AND*. Los usuarios necesitan combinar sinónimos y cuasi sinónimos de distintos conceptos en una misma necesidad de información, y para hacer eso se necesitan *OR*s y *AND*s, y a veces es preciso excluir cosas del resultado, y hace falta entonces un *NOT*, etc.

La cuestión es que *Google* ha acabado montando una opción de búsqueda avanzada con otras opciones además del *AND*.

Volviendo al intento de hacer una prospectiva, lo que puedo decir es lo siguiente: si algún día se diseña un sistema de búsqueda avanzada, será porque detrás del mismo habrá un sistema experto capaz de convertir una frase de lenguaje natural en una ecuación booleana...

***¿Podemos esperar que pronto haya un Google para la internet invisible?***

—Sí. Si entendemos por internet invisible los contenidos que interesan a la mayoría de los documentalistas, es decir, publicaciones en formatos no html y contenido de bases de datos no indizables hasta ahora por motores de búsqueda, la respuesta es un rotundo sí. Costará más o menos y serán sistemas más o menos completos, pero los tendremos. El propio *Google* ya ha hecho retroceder un poco las fronteras de la internet

invisible al indizar documentos pdf y de otros formatos; y tenemos bases de datos gratuitas que indizan los contenidos de revistas digitales. Hay algunas sedes web que también afirman hacer búsquedas en el interior de bases de datos, como *PlanetSearch*. Por ahora son casos muy limitados, pero irán creciendo.

***¿Cree que la web semántica es una realidad a corto plazo o aún tendremos que esperar muchos años?***

—Todo parece indicar que es algo a medio y largo plazo, de ninguna manera a corto plazo. ¿La razón? Por ahora depende de la simple voluntad de editores y de creadores de páginas web. No hay ningún aliciente (y eso es malo) ni ninguna coerción (y eso es bueno) para crear sedes web y para publicar información digital en los formatos que se suponen que favorecen la web semántica (xml, *Dublin Core*, *RDF*, etc.), por tanto tenemos para rato. Siempre hay excepciones, como es lógico en algo tan vasto como la web. Pero, si no hay algún cambio radical en los editores de páginas web, los servidores web, las normas actuales de publicación, etc., la inmensa mayoría de las páginas web estarán editadas en código "no semántico" como el html puro y duro o como código binario (ejemplo, *Flash*) que se activa con una llamada desde una etiqueta html, etc.

***¿Puede indicar tres tendencias de futuro en la búsqueda de información en la web para los próximos cinco años?***

—La primera: el audiovisual. La web será cada vez más audiovisual, sin perder su carácter textual.

La segunda: el multimedia, que es el paso lógico después del audiovisual o en paralelo a aquél cuando el audiovisual está en soporte digital. Es decir, aparecerán formas de interactividad nuevas, más allá del simple *click* de enlace a enlace. Aparecerá la interactividad propia del multimedia: gráficos interactivos y simulaciones. Un adelanto lo tenemos en las secciones multimedia de diarios en línea como *El País*, *El Mundo* o *New York Times*.

La tercera: el auge de la arquitectura de la información y de los sistemas de información con meta información en el diseño de páginas web, pero no en el sentido de *Dublin Core*, sino en el sentido documental de siempre: sistemas donde la información se estructurará en campos y donde la metainformación en forma de categorías, taxonomías, descriptores o palabras clave será esencial. Esta tercera tendencia se limitará, seguramente, al campo de las sedes webs relacionadas con la cultura, la ciencia, el pensamiento y el sector público en general.

provecho de las relaciones semánticas entre conceptos de interrogación y contenidos.

—Incorporación de módulos que permiten la generación de resúmenes teniendo en cuenta la distancia entre las diferentes partes del texto, para lo cual utilizan "diccionarios geográficos".

—Experimentación de métodos para combinar diversas cadenas léxicas.

A iniciativa de la *Universidad Carlos III*, en estas jornadas se presentó la *CDU* como un novedoso instrumento para la **clasificación** automática partiendo de la idea de procedimiento de **Ranganathan**. A partir de un índice alfabético de materias y términos del lenguaje natural asociados a cada notación numérica de la *CDU*, se le vinculan términos del lenguaje natural así como encabezamientos de materia.

**«Se destacó el papel de los agentes inteligentes como sistemas que proponen soluciones importantes tales como la eficacia, la interactividad y la movilidad en la recuperación de la información»**

Por otra parte, ante la falta de avances en el campo de la **indización** clásica (descriptores o palabras clave, por ejemplo) se propone la indización conceptual o semántica como solución, por un lado, a la relación coste-beneficio de la anterior y, por otro, a los problemas de desambiguación (que no termina de solventar).

El procedimiento automatizado de indización se ha perfeccionado gracias al análisis morfosintáctico y semántico de las técnicas manuales. A diferencia de éste, la mayoría de procesos de análisis lingüístico (como la lematización, que consiste en la búsqueda de la palabra en su forma básica) no produce mejoras claras y sí supone un coste adicional de procesamiento muy notable a causa de la sinonimia y la polisemia.

Los **mapas conceptuales** se conciben como herramientas que gestionan el conocimiento y que organizan la representación de conceptos. Se perciben como una herramienta didáctica útil para promover la adquisición de conocimiento.

Los ***topic maps***, por su parte, son documentos o conjuntos de documentos sgml o xml interrelacio-

nados en un espacio multidimensional en el que cada una de sus partes (localizaciones) se denominan *topics*.

En referencia a los **tesauros**, otra de las facetas que se constató fue su capacidad de innovación en el campo de las tecnologías de la información.

Durante las jornadas se presentaron los mapas conceptuales, los tesauros y los *topics maps* como tres formas complementarias de organizar y representar la información, ya que cubren aspectos distintos en su búsqueda y recuperación.

Una de las técnicas que más incidencia está teniendo en esta última ha sido la aplicación de las **redes neuronales** en lo que se refiere a la comparación de parejas de términos con el fin de saber si son multipalabras (sucesión de términos cuyo significado es diferente a la suma de dichos términos, por ejemplo "casa blanca"). Para ello se han desarrollado redes (neuronales y bayesianas) que trabajan en la clasificación de multipalabras exógenas. Los experimentos han demostrado que ambos métodos mejoran la precisión alcanzada por un sistema de recuperación de información.

**2. Recuperación de la información**

En el área de la **cibermetría** se trató el problema que presenta la información estructurada a la hora de recuperarla en la web, para lo cual se vienen utilizando agentes software. La línea de investigación que se ha desarrollado consigue:

—Recuperar información en la web.

—Niveles de precisión muy buenos.

—Reducir el número de nodos o visitas.

Este sistema se plantea como una solución muy interesante para la recuperación de información multilingüe.

**«El camino hacia el tratamiento y la recuperación de información está en la aplicación de la indización conceptual frente a las técnicas lingüísticas clásicas»**

El desarrollo de los agentes software en la creación de algoritmos y de estudio de técnicas cibermétricas intenta llegar a reducir el número de nodos a recorrer para obtener altos niveles de precisión en la respuesta.

Los **directorios web** (que se presentaron como modalidad de los buscadores) son taxonomías que clasifican documentos web sobre los que posteriormente se realizan consultas.

Tales sistemas permiten un tipo determinado de búsquedas en las que la colección de documentos permanece restringida a una zona de las categorías.

Esta arquitectura se basa en una estructura de datos híbrida constituida por un fichero invertido que contiene ficheros de firmas. Éstos consisten en un sistema de indización que contrasta un término a través de una categoría. Sobre la arquitectura se definen dos modelos: el modelo híbrido con información total (que contrasta todo el contenido del documento) y el modelo híbrido con información parcial (que contrasta partes del documento).

La búsqueda está restringida sólo a los documentos que pertenezcan a una categoría en concreto, limitando la colección que se ha de buscar.

El problema de los ficheros de firmas es que suelen tener falsos aciertos al confundir los términos.

La conclusión a la que se llegó fue que resulta más práctico procesar a través del modelo híbrido de información parcial porque optimiza el tiempo de respuesta al tratar sólo partes del documento y no el texto completo.

La inteligencia artificial es una línea transversal de investigación entre la recuperación de información y la representación del conocimiento. Desde este área los **agentes inteligentes** son algo innovador en la organización de la información. Durante las jornadas se destacó su papel en la propuesta de soluciones de cara a una mayor eficacia, interactividad y movilidad en la respuesta.

Por su parte, los **algoritmos** están constituidos por una serie de pasos organizados que describen un proceso a seguir para dar solución a un problema específico. Se aplican en la indización de documentos, agrupamiento de documentos y términos, definición de consultas, cálculo de la relevancia, etc.

En este punto se trataron los algoritmos genéticos, que presentan grandes avances al intentar aprender las estructuras de los términos, así como los pesos y conexiones entre términos por medio de diferentes técnicas.

El **algoritmo de Kohonen** es utilizado para localizar las relaciones contextuales existentes entre distintos términos presentes en un corpus documental. Realiza agrupaciones (*clusters*) y establece una organización topológica de los mismos. De ahí que su aplicación en la recuperación de la información pueda ser muy útil, ampliando las búsquedas a través de términos relacionados con el que se ha iniciado la consulta.

**—Casos y experiencias**

Además de las líneas de trabajo e investigación que se mostraron

en las diferentes ponencias y comunicaciones, en *Jotri'2002* se presentaron también los resultados de diversas experiencias y prácticas desarrolladas. De todas ellas se exponen a continuación algunos ejemplos:

**Proyecto *Hermes***

En *Hemerotecas electrónicas: recuperación multilingüe y extracción semántica* participan la *Univ. Nacional de Educación a Distancia* (*Uned*), la *Universidad del País Vasco* (*EHU*) y la *Universitat Politècnica de Catalunya* (*UPC*). Uno de sus propósitos es la confección de resúmenes automáticos de noticias periodísticas.

Para poder evaluar los sistemas de resumen automático se dispone de un corpus de evaluación coherente, compuesto por textos originales (inicialmente un número de noticias de la agencia *EFE* en lengua castellana) con sus correspondientes resúmenes realizados por personas.

El objetivo es que, una vez analizados los resultados, se complete el corpus con otros documentos y para todas las lenguas implicadas.

*http://terral.lsi.uned.es/hermes/*

**Modelo de recuperación de información multilingüe.**

*Cross-language information retrieval* (*Clir*)

Ante los problemas que suponen las barreras lingüísticas en los sistemas de recuperación de información en las colecciones multilingües, se presentaron estudios sobre este campo proponiendo soluciones como los sistemas de traducción automática (en tanto que respuesta a los problemas de traducción de la consulta) y los tesauros de similitud multilingües (en los casos de los corpus comparables).

Los sistemas utilizados en la traducción de la consulta del usuario o de los documentos presentan problemas como la disponibilidad para idiomas no muy extendidos y la dependencia de la lengua origen y destino en la calidad del resultado.

Los corpus comparables no requieren que un documento sea traducción de otro, sino que dado un documento existan textos en otros idiomas que traten el mismo tema.

Tanto los sistemas de traducción automática como los tesauros de similitud multilingües se experimentan en entornos web, y apuntan una tendencia hacia un alto grado de independencia del idioma, centrándose en los recursos lingüísticos.

*http://www-csli.stanford.edu/semlab/infomap/CLIR.html*

***Citec***

*Citec* (originariamente *Citas de economía*) es un agente software creado para llevar a cabo enlaces de referencias existentes en la biblioteca digital. Trabaja en el entorno de los documentos científicos a través de citas y referencias bibliográficas, que se utilizan como herramienta para la recuperación de información. Busca en un texto la sección donde están las referencias, examina en la red si se encuentran estos documentos referenciados y, de ser así, crea los vínculos entre ambos.

*Citec* se basa en unidades de metadatos y no en el texto completo de los documentos; de esta manera los pasa de pdf a ascii para que el motor encuentre todos los documentos que son referenciados en un documento dado. Si los tiene la biblioteca digitalizada entonces crea un enlace entre los que son citados y el que cita.

A través de este agente software, se realiza la extracción y enlace automáticos de referencias bibliográficas en documentos electrónicos.

*http://netec.ier.hit-u.ac.jp/CitEc/*

## Conclusiones

Durante estas jornadas se ha coincidido en que el camino hacia el tratamiento y la recuperación de información está en la aplicación de la indización conceptual frente a las técnicas lingüísticas clásicas de indización.

Por otra parte, es de recibo reconocer la correcta organización del evento y lo acertado de su celebración. Sin embargo, se ha advertido en *Jotri' 2002* un exceso de enfoque informático, que en ocasiones ha llegado a programación pura, haciéndose incomprensible a otros colectivos ajenos. En cambio desde la biblioteconomía y la documentación se han limitado las intervenciones a la descripción de experiencias de desarrollo de resúmenes, índices o tesauros.

La mayoría de casos expuestos eran experimentos de laboratorio, estudios que se han llevado a cabo en una población cerrada en lugar de seleccionar una muestra relevante de la población de la Red, con lo que los resultados pueden resultar insuficientes a la hora de aplicarlos a internet.

Esperemos que en la segunda edición de *Jotri* (como en otros acontecimientos interdisciplinares de este tipo, como *Jbidi*, por ejemplo) la presencia y el papel de los profesionales de la información se haga más palpable.

*http://www.fiv.upv.es/jotri/principal.htm*

***Susana Serrano González***

*susana.serrano@gtd.es*

***José Antonio Ontalba y Ruipérez***

*jontalba@uoc.edu*