

SOFTWARE DOCUMENTAL



CARROT2: BÚSQUEDA Y VISUALIZACIÓN DE LA INFORMACIÓN



Audilio Gonzales-Aguilar y María Ramírez-Posada



Audilio Gonzales-Aguilar es doctor en derecho y nuevas tecnologías por la *Universidad de Montpellier I*. Es profesor titular de la *Universidad Paul Valéry Montpellier 3* en Montpellier, Francia, donde es miembro del equipo de investigación *Praxiling UMR 5267 CNRS—Université Paul-Valéry—Montpellier 3*. Director del master profesional en gestión de la información y de la documentación. Sus áreas de investigación son el análisis de redes sociales aplicadas a contenidos documentales, hipertextualidad del documento digital y visualización y representación de la información

Universidad Paul Valéry Montpellier 3

Departamento de Documentación

Route de Mende, 34199 Montpellier Cedex 5, Francia

<http://www.univ-montp3.fr>

audilio.gonzales@univ-montp3.fr



María Ramírez-Posada es bióloga marina, candidata a maestría en desarrollo sostenible y medio ambiente (*Universidad de Manizales*, Colombia) y candidata a maestría en gestión de la información y del conocimiento (*Universidad Paul Valéry Montpellier 3*, Francia). Cuenta con experiencia como investigadora en el área del desarrollo sostenible empresarial y visualización de la sustentabilidad y como docente a nivel pre gradual y pos gradual.

Grupo de Investigación Eco-ambiental

<http://201.234.78.173:8080/gruplac/jsp/visualiza/visualizagr.jsp?nro=0000000005056>

webdocumenta@gmail.com

Resumen

Obtener información en cortos períodos de tiempo puede marcar la diferencia en caso de una investigación o de un negocio, por lo cual es importante contar con herramientas que faciliten su búsqueda. Se analiza una técnica de visualización conocida como motor de agrupamiento de información: *Carrot2*, que separa los documentos –tanto encontrados en la Web como por grupos o clústeres, utilizando algoritmos de agrupamiento que, mediante la consulta de diversas fuentes de datos, los procesa y muestra la visualización de la información obtenida. Este software de código abierto reduce significativamente el esfuerzo de la recuperación de la información y su análisis, organizándola por grupos temáticos.

Palabras clave

Visualización de la información, Recuperación de información, Búsqueda web, *Clustering*, Árbol de resultados.

Title: *Carrot2*: Information search and visualization

Abstract

Obtaining information in short periods of time can make the difference for research or business; therefore it is important to have tools that facilitate the search for information. A visualization technique called information-clustering engine is reviewed: *Carrot2* separates the documents found on the Web into groups or clusters, using clustering algorithms that consult various sources of data and then process and display the information obtained. This open-source software significantly reduces the effort involved in information retrieval (IR) and analysis.

Keywords

Information visualization, Information retrieval, Web search, Clustering, Results tree.

Gonzales-Aguilar, Audilio; Ramírez-Posada, María. “*Carrot2*: búsqueda y visualización de la información”. *El profesional de la información*, 2012, enero-febrero, v. 21, n. 1, pp. 105-112.

<http://dx.doi.org/10.3145/epi.2012.ene.14>

Artículo recibido el 11-01-12

Aceptación definitiva: 18-01-12

Introducción

Con el avance de internet y la existencia de nuevas tecnologías y servicios, el acceso a la información ha tenido un crecimiento sin precedentes. Al buscar información, lo que normalmente se hace es explorar, describir y organizar manualmente los resultados recuperados¹. Pero hoy en día se pueden usar tecnologías eficientes y flexibles, capaces de combinar la recuperación y la posterior organización de la información.

Los motores de búsqueda muestran una lista de resultados por orden de relevancia, un ranking, y el usuario tiene que examinarla de forma descendente, hasta hallar la información solicitada. No hay forma de determinar exactamente qué es relevante para el usuario, ya que las consultas suelen ser muy breves y su interpretación es confusa en ausencia de un contexto. Por ello, a pesar de que los motores de búsqueda son buenos para ciertas tareas, pueden ser menos efectivos para satisfacer determinadas demandas amplias o, al contrario, dar respuesta a preguntas definidas.

El acceso a la información ha tenido un crecimiento sin precedentes con el uso de nuevas tecnologías y servicios

Un enfoque diferente es el de la agrupación de los resultados en los denominados clústeres. El programa *Carrot2* realiza el *clustering*² (categorización o agrupamiento) de los documentos hallados basándose en similitudes entre ellos, sin un conocimiento a priori de sus características (**Goldenberg, 2007**), lo cual permite mejorar la precisión. El usuario

plantea un tema general y posteriormente puede pasar a analizar los temas más específicos creados de forma dinámica a partir de los resultados de la consulta.

<http://project.carrot2.org>

Carrot2 facilita ampliar o modificar la estructura de la búsqueda, comprender mejor tema, y favorece la exploración sistemática de los resultados.

La agrupación de documentos en clústeres o grupos permite mejorar la precisión de la recuperación de información

Características

Carrot2 es un software de recuperación, *clustering* y visualización de documentos y contenidos web. Contiene una colección de algoritmos³ de agrupación que facilitan la exploración del contexto temático de los documentos recuperados a través de los motores de búsqueda o de una colección de textos en un ordenador o en un servidor⁴.

Instalación

Carrot2 puede ser utilizado online como metamotor o buscador de contenidos web a partir de cualquier navegador. Es posible integrar directamente los plug-in, o extensiones, para *Firefox* e *Internet Explorer*.

<http://search.carrot2.org/stable/search>

<http://project.carrot2.org/download-search-plugins.html>

En su versión de escritorio local se puede descargar para los sistemas *Windows*, *Mac OS* y *Linux*. Asimismo, presen-

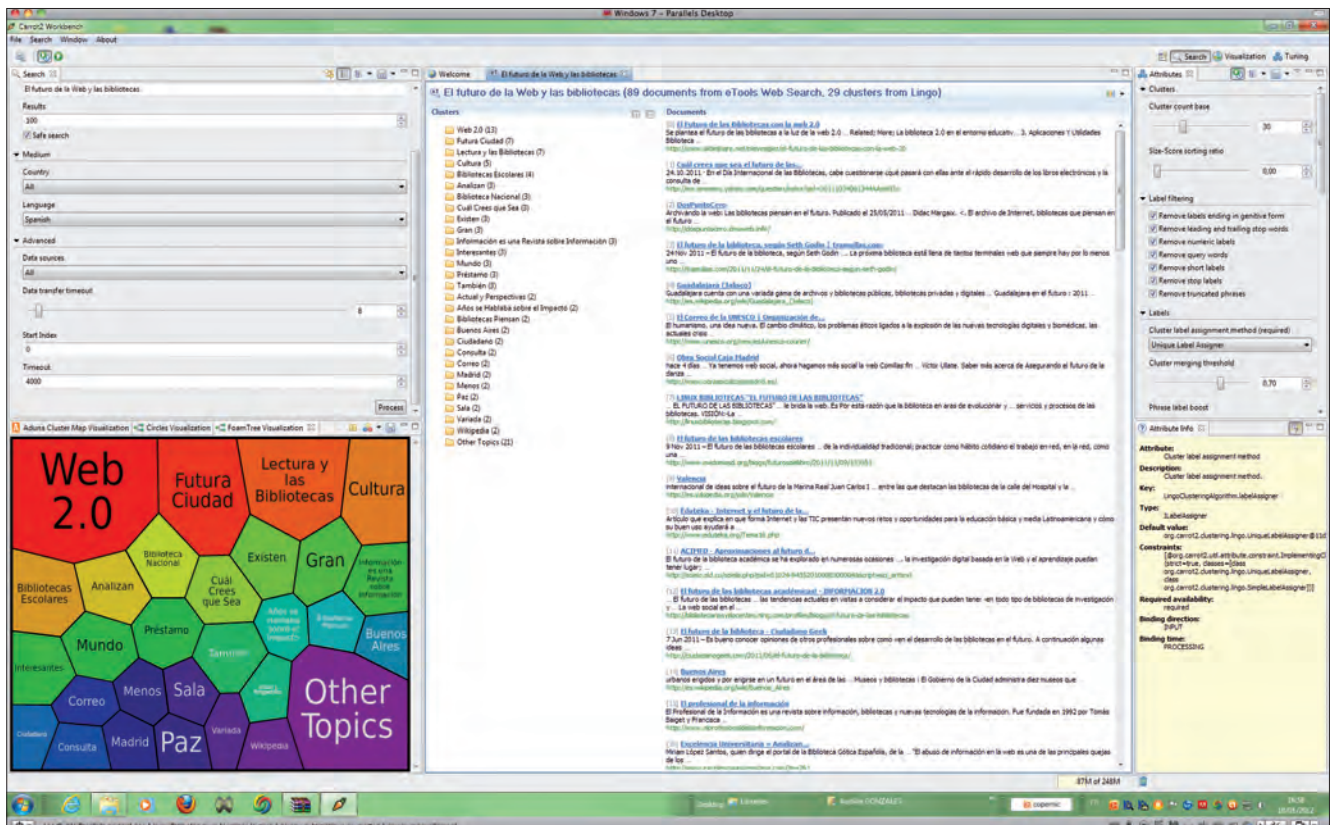


Figura 1. *Carrot2* crea clústeres de etiquetas de un conjunto de documentos.

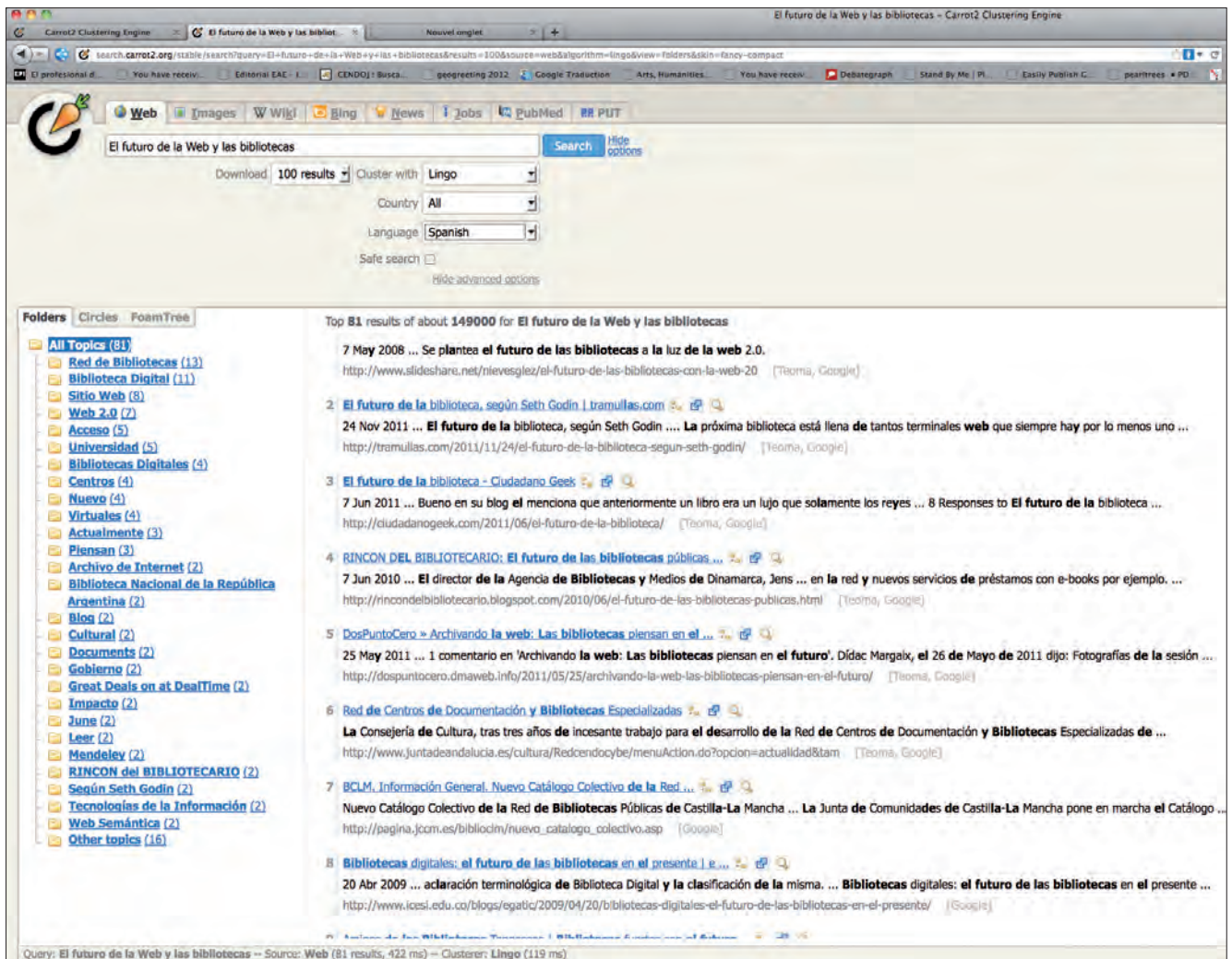


Figura2. Interfaz de Carrot2 en línea

ta otras posibilidades de descarga y utilización: en formato java, integración con PHP, C# y además se puede instalar como una instancia de búsqueda en un sitio web.

<http://project.carrot2.org/download-workbench-win32.html>

<http://project.carrot2.org/download-workbench-macos-cocoa-32bit.html>

<http://project.carrot2.org/download-workbench-linux.html>

<http://project.carrot2.org/download.html>

Carrot2 es un software de recuperación, clustering y visualización de documentos que se puede usar tanto online en su web como en nuestro ordenador

En Windows está disponible para win32 y para win64 (sistemas basados en una arquitectura de 32 ó 63 bits). Una vez descargado, se descomprime y se ejecuta el archivo *carrot2-workbench.exe*. El programa no requiere ninguna instalación adicional. Además, si el ordenador personal tiene instalado *Google Desktop*, Carrot2 recupera toda la información indexada con este medio, e igualmente puede emplearse para la recuperación de documentación dentro de un equipo cliente.

<http://googledesktop.blogspot.com>

Interfaz de Carrot2

Carrot2 tiene una interfaz que corresponde a tres fases del proceso de la información: entrada, filtrado y salida de la información. En la figura 3 se pueden observar los componentes de la aplicación de escritorio y en la figura 4 los de la interfaz en línea.

Opciones de búsqueda

Comprende:

- Fuentes de información, referentes a los motores de búsqueda como *Google*, *Yahoo*, *Wiki*, *PubMed*, etc.
- Algoritmos para la creación de grupos o clústeres como *Lingo* (que se emplea por defecto), *K-mean*, *STC*. También puede agrupar por dominio (.net, .com, .org...).

Páginas de resultados

Muestra un listado de todos los clústeres identificados y de los documentos de cada clúster por orden de relevancia.

Formas de visualización gráfica

Carrot2 presenta tres tipos de representación gráfica, como se observa en la figura 6:

- Esquema relacional *Aduna*: visualiza las relaciones entre los clústeres. Desarrollado por <http://www.aduna-software.com>

- Diagrama circular.
- Mapa de áreas, utilizando colores cálidos para las más pertinentes y colores fríos para las menos en la parte inferior. En la base del mapa aparece el grupo desconectado "Other topics".

Configuración de atributos

En la versión de escritorio *Carrot2* integra una ventana de atributos en la que se pueden configurar y editar las dimensiones de agrupamientos. Esta ventana se encuentra a la derecha de la pantalla del programa.

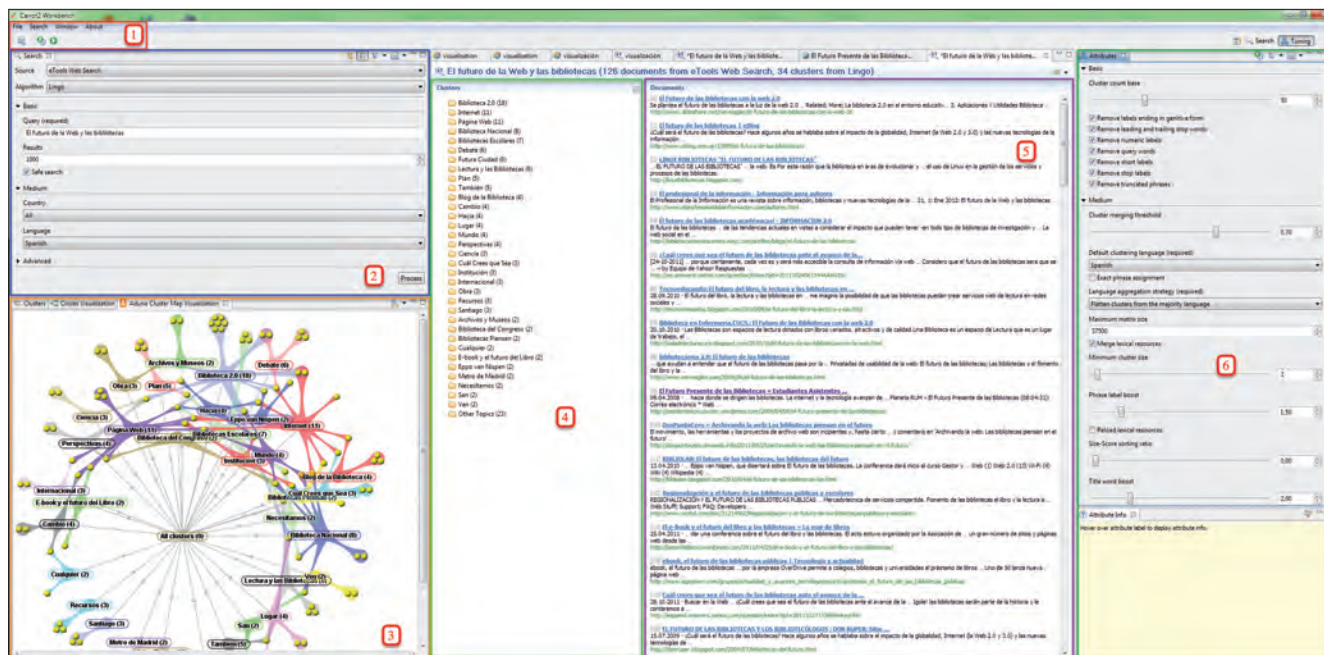


Figura3. Ventanas de la interfaz *Carrot2* en la aplicación escritorio.
1. Menú del programa, 2. Opciones de búsqueda, 3. Visualización, 4. Ventana de clústeres, 5. Resultados, 6. Edición y atributos de la consulta.

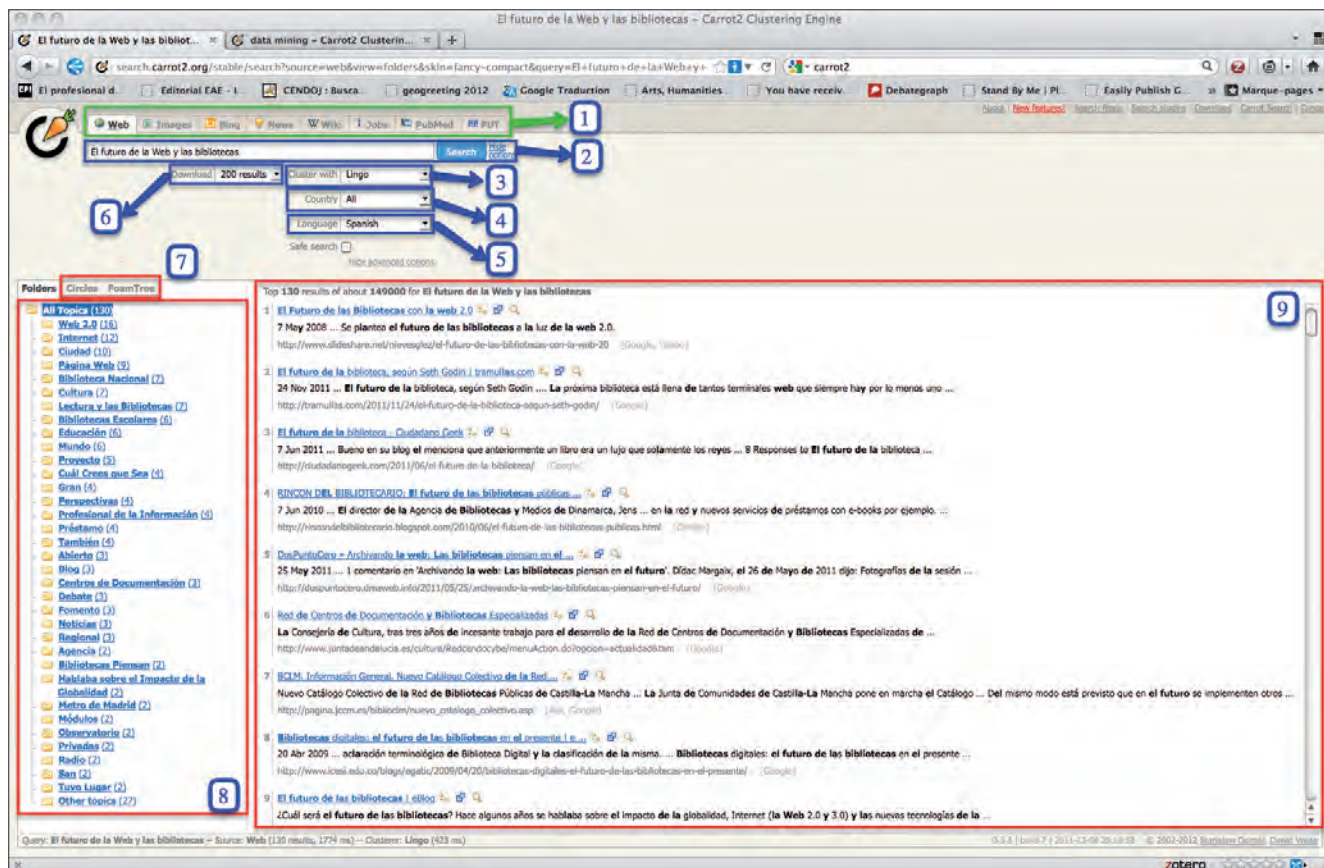


Figura 4. Elementos que componen la interfaz de *Carrot2* en línea.
1. Opciones de búsqueda, 2. Campo de búsqueda, 3. Ventana de clústeres, 4. País, 5. Idioma, 6. Numero de resultados solicitados, 7. Visualización (diagrama circular, mapa de superficie por grupos), 8. Ventana de clústeres, 9. Resultados.

Conclusión

Es de destacar la flexibilidad de esta aplicación, que permite hacer una *radiografía* de los resultados de una búsqueda de información documental. Cuando se señala que haga una búsqueda en la Web en general (opción *eTools Web Search*) la hace simultáneamente en *Ask*, *Bing*, *Entireweb*, *Teoma*, *Yahoo*, *Google* y *Wikipedia* y forma un conjunto con los primeros resultados de cada buscador. También se pueden especificar determinados espacios de búsqueda como imágenes, wikis, noticias, arte, *PubMed*, *Google Desktop* (sobre los documentos de nuestra computadora), etc. Estaría bien que en el futuro añadieran *Google Scholar*.

Flexibilidad, pertinencia y excelentes opciones gráficas son características básicas de *Carrot2*

Al usar el programa, la lista de los primeros 100 resultados (la cantidad puede variarse) se representa con un árbol clasificándolos a partir de las palabras que figuran en los documentos. Como fallo del sistema –y de todos los buscadores agrupadores– hay que indicar que algunos de los conjuntos se basan en palabras vacías (por ejemplo, en la figura 6 vemos “basados”, “existen”, “habló”, “también”, “cualquier”...), lo cual disminuye mucho su eficacia.

Clasificación jerárquica de los resultados, con una interfaz fácil de trabajar

Sin embargo, a la falta de precisión de las búsquedas, como ocurre con todos los sistemas de recuperación de información, hay que encontrarle las ventajas de la serendipidad: hallar cosas interesantes sin buscarlas específicamente, y aquí con una representación gráfica, al ser los conceptos más evidentes, el descubrimiento es más fácil.

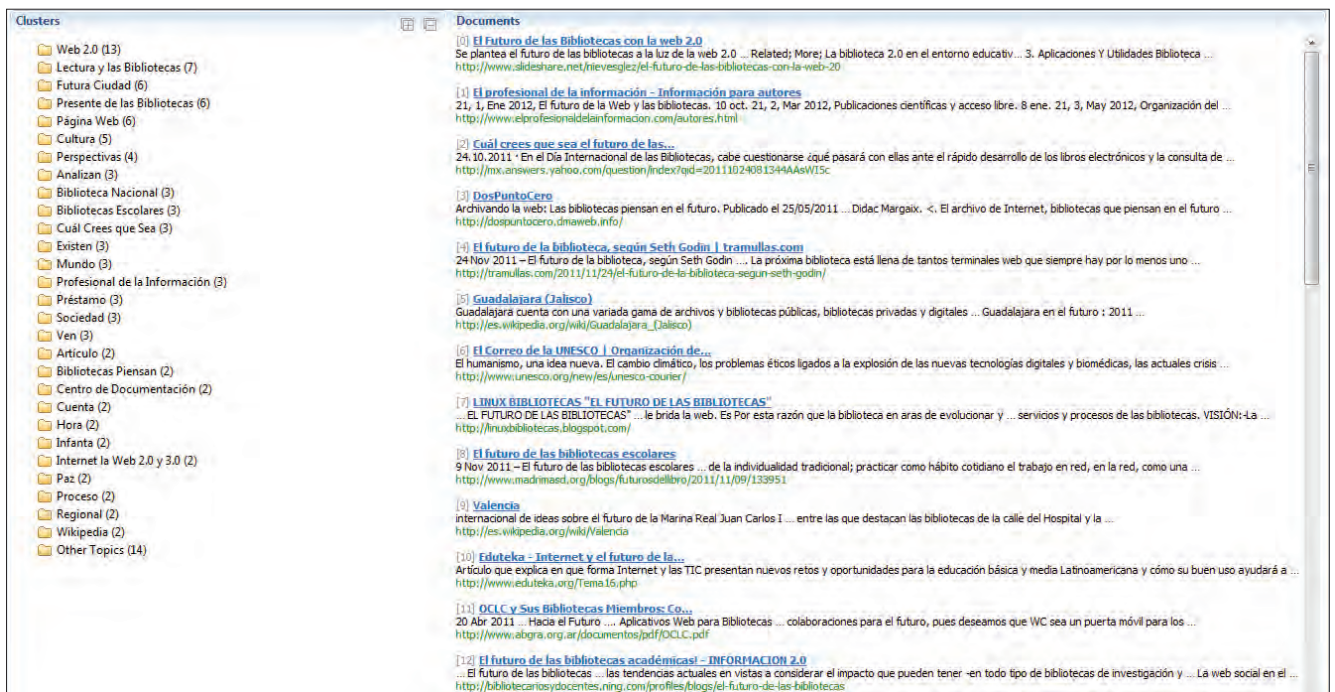


Figura 5. Ventana de resultados (clústeres y documentos)



Figura 6. Formas de visualización gráfica de *Carrot2*

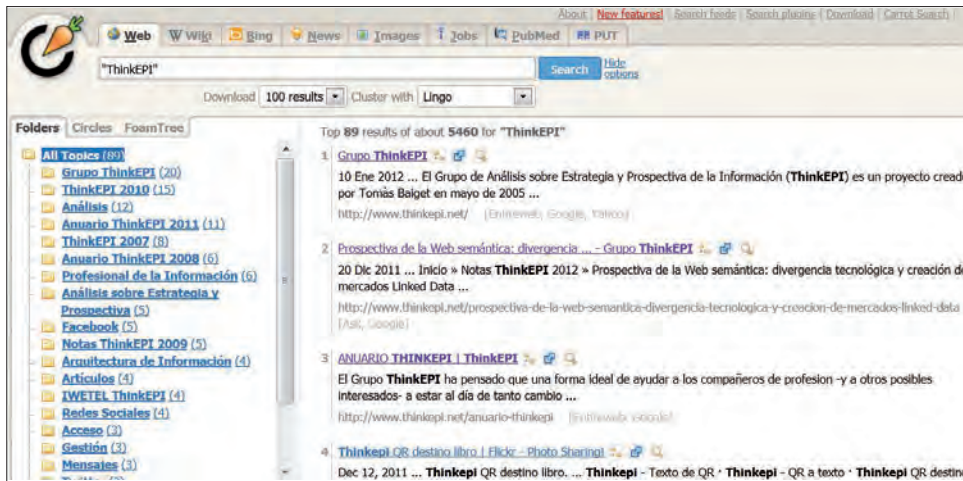


Figura 7. Carrot2 da más precisión añadiendo comillas. Es conveniente ponerlas incluso con palabras sueltas, pues el sistema las descompone. En este ejemplo buscaría también “think” y “epi”

Al hacer clic en cualquiera de los grupos o clúster se despliega una lista de los resultados correspondientes.

La visualización de la información en Carrot2 permite la clasificación de los resultados de búsqueda, la optimización de los mismos y la extracción de palabras clave.

Notas

1. **Dürsteler, Juan C.** Visualización de la información: proceso de interiorización del conocimiento mediante la percepción de información.

<http://www.infovis.net/printMag.php?num=100&lang=1>

2. El término *clustering* o agrupación hace referencia a aquellos sistemas de recuperación que emplean algoritmos de agrupación de contenidos, y cumplen la tarea de particionar un espacio de información no etiquetada en grupos, clases o clústeres.

3. Carrot2 organiza los documentos y contenidos web utilizando algoritmos de agrupación. Lingo es uno de ellos, que logra bastante calidad en etiquetar los grupos basándose en frases recurrentes y crea un árbol de resultados.

4. Los ejemplos típicos para demostrar esta técnica muestran el contexto de las consultas amplias y ambiguas de sinonimia conceptual como “apache” (helicóptero, tribu indígena o software) o “salsa” (danza o comida).

Bibliografía

Anastasiu, David C.; Gao, Byron J.; Buttler, David. “ClusteringWiki: a framework for personalized clustering of search results”. En: *Sigir’11 Procs of the 34th intl ACM Sigir conf on research and development in information retrieval*, 2011. <http://dmlab.cs.txstate.edu/ClusteringWiki/pdf/cw.pdf>

Carpineto, Claudio; Osinski, Stanislaw; Romano, Giovanni; Weiss, Dawid. “A survey of web clustering engines”. *ACM computing surveys*, 2009, July, v. 41, n. 3, pp. 17-38. <http://search.fub.it/claudio/pdf/CSUR09.pdf> <http://dx.doi.org/10.1145/1541880.1541884>

Cigarrán-Recuero, Juan-Manuel. *Organización de resultados de búsqueda mediante análisis formal de conceptos.*

Tesis doctoral. Universidad Nacional de Educación a Distancia, 2008.

<http://e-spacio.uned.es/fez/view.php?id=tesisuned:IngInf-Jcigarran>

Goldenberg, Daniel. *Categorización automática de documentos con mapas auto-organizados de Kohonen.* Tesis de magister. ITBA, 2007. <http://www.itba.edu.ar/archivos/secciones/goldenberg-tesisdemagister.pdf>

Laufert, Thomas. “Visualization of the Carrot 2 system”. Baltimore: University of Mary

land Baltimore County.

<http://www.csee.umbc.edu/conference/src/conferences/src1/www/papers/laufert.pdf>

Osiński, Stanislaw. “Improving quality of search results clustering with approximate matrix factorisations”. En: *Lecture notes in computer science, Procs of the 28th European conf on IR research*. 2006, v. 3936, pp. 167-178. http://dx.doi.org/10.1007/11735106_16

Osiński, Stanislaw; Weiss, Dawid. “Carrot2: design of a flexible and efficient Web information retrieval framework”. En: *Lecture notes in computer science. Procs of the Third intl Atlantic web intelligence conf (AWIC 2005)*, Łódź, Poland, 2005, v. 3528, pp. 439-444. <http://www.cs.put.poznan.pl/dweiss/site/publications/download/dweiss-carrot2-poster-awic2005.pdf> http://dx.doi.org/10.1007/11495772_68

Osinski, Stanislaw; Weiss, Dawid. *Clustering search results with Carrot2.* Polonia: Poznan Supercomputing and Networking Center; Institute of Computing Science, Poznan University of Technology, 2007. <http://project.carrot2.org/publications/carrot2-dresden-2007.pdf>

Senthil-Kumar, R. Subhashini; Senthil-Kumar, V. Jawahar. “The anatomy of web search result clustering and search engines”. *Indian journal of computer science and engineering*, 2010, v. 1, n. 4, pp. 392-401.

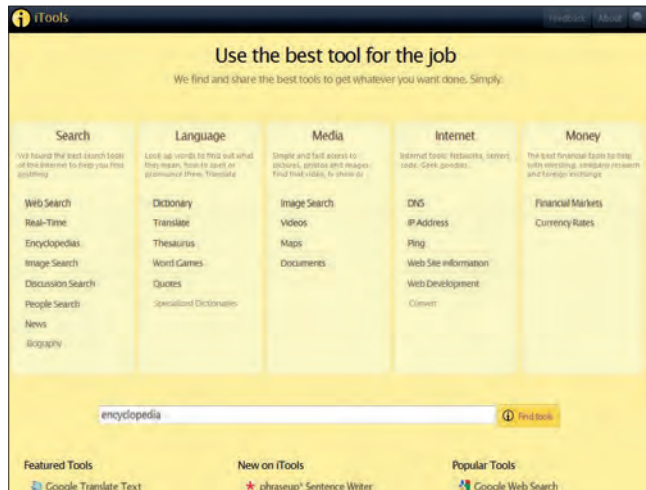
Stefanowski, Jerzy; Weiss, Dawid. “Carrot2 and language properties in Web search results clustering”. *Lecture notes in artificial intelligence: advances in Web intelligence. Procs of the First intl Atlantic web intelligence conf*, Madrid, Spain, 2003, n. 2663, pp. 240-249. http://dx.doi.org/10.1007/3-540-44831-4_25

Weiss, Dawid; Osinski, Stanislaw. “Carrot2 clustering framework”. *Procs of the First intl conf on open source systems*, 2009, pp. 298-299.

Weiss, Dawid. *Descriptive clustering as a method for exploring text collections.* Tesis doctoral. Institute of Computing Science, Poznan University of Technology, 2006. <http://www.cs.put.poznan.pl/dweiss/site/publications/download/dweiss-phd-thesis.pdf>

iTools

<http://itools.com>



Es un “buscador de herramientas de la Web”. En el buscador inicial se pregunta el tipo de información o servicio que se desea (en inglés): enciclopedias, traductores, vídeos, idiomas, finanzas..., y el sistema ofrece diferentes opciones donde buscar. Además es como un portal con multitud de recursos agrupados por tipos, que se abren y se usan haciendo clic en ellos sin más.

Yippy

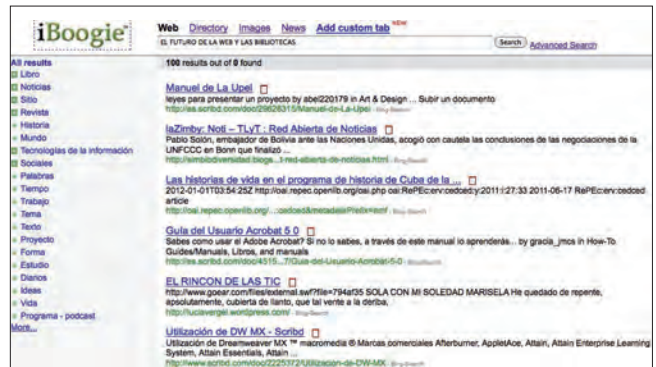
<http://www.yippy.com>



Metabuscaador bastante estándar, como los ya descritos anteriormente. Tiene una pestaña llamada *Yippy labs*, con aplicaciones concretas como un buscador adaptado para ser usado desde la *Wii* de *Nintendo*, nube de etiquetas, buscador sobre Shakespeare, y buscador sobre Benjamin Franklin. Tiene un filtro para contenidos no apropiados para niños.

iBoogie

<http://www.iboogie.com>



Las etiquetas de los grupos que forma aparecen a la izquierda. Su característica distintiva es *Custom tab*, que permite añadir webs concretas pre-seleccionadas de Blogs, Alimentos, Juegos, Gobierno de los EUA, Editorial IDG, Israel, Bibliotecas, Medicina, Buscadores, Noticias, etc.

Hakia

<http://www.hakia.com>



Uno de los mejores buscadores de esta lista. Organiza los resultados por diferentes tipos de fuentes de información, incluida la “credibilidad”, que contiene los sitios recomendados. Se presenta como *buscador semántico* sobre medicina (ha realizado <http://newpubmed.com> para *PubMed*), finanzas para inversores (*MoodTrade*), industria aeronáutica (*AeroHakia*). Dice no sólo “leer” las palabras, sino que las “interpreta”.