UNIVERSIDAD POLITÉCNICA DE VALENCIA

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

LICENCIATURA EN DOCUMENTACIÓN

Conducta de búsqueda de usuarios de internet en *google.com*

Proyecto Final de Carrera presentado por:

ANA CASADO SÁNCHEZ

Directora:

NURIA LLORET ROMERO

Codirector:

JORGE SERRANO COBOS

Valencia, septiembre de 2011

ÍNDICE

ÍNDICE DE FIGURAS

1. INTRODUCCIÓN

Internet ha supuesto, desde su popularización a mediados de los años 90, una revolución en el ámbito del acceso a la información. Antes de la llegada de Internet, la biblioteca era *la protagonista del mundo de la información* (Lozano, 2010). En la actualidad los buscadores se han convertido en *la primera herramienta de búsqueda de información y servicios para el 80% de los usuarios de internet* (Jansen, 2004).

La adaptación por parte del usuario parece que ha sido rápida. Los primeros buscadores de la Web emulaban las funcionalidades de los buscadores tradicionales utilizadas en los OPAC (Catálogos en línea de las bibliotecas). Daban la posibilidad de realizar búsquedas complejas utilizando operadores *booleanos* (AND, OR, AND NOT,…), búsquedas por campos,… pero algunos estudios reflejan en sus resultados que menos del 5% de las búsquedas en internet recogidas utilizaba operadores booleanos (Spink, Wolfran, Jansen y Saracevic, 2001) confirmando que el usuario no estaba interesado en utilizar la tradicional *búsqueda avanzada.*

Además de los buscadores, la Web 2.0, entendida como *plataforma* (O'Reilly, 2006), proporciona una serie de aplicaciones, blogs, wikis, redes sociales, entornos para compartir recursos, en las que el usuario aporta y comparte conocimientos e información con el resto de usuarios y ha generado un nuevo escenario en el que *las interfaces de usuario y los sistemas de información se han hecho cada vez más sociales, destinados a apoyar la producción descentralizada, de cooperación y el uso de contenidos. (Pirolli, 2009)*.

Y el cambio va a más. La tecnología avanza y sigue modificando nuestros hábitos. Nuevos dispositivos favorecen nuevas conductas. Los dispositivos móviles con acceso a Internet permiten estar conectados a la red en cualquier lugar, las 24 horas al día. El usuario maneja un gran volumen de información, y tiene la posibilidad de compartirla con otros usuarios y retroalimentarse.

> *La demanda de la banda ancha en España sigue creciendo. Durante 2010 se han sobrepasado los 10 millones de líneas de banda ancha fija habiéndose experimentado un crecimiento del 8,3% en el último año. Por otro lado, la realidad móvil ha crecido aún más: Telefónica cuenta con más de 4,3 millones de usuarios de banda ancha móvil, 2 veces más que en el año 2009, lo que supone un crecimiento interanual del 58,7%. En el caso de los datacards el crecimiento ha sido de casi 900.000 unidades, lo que supone un ascenso en el último año de un 60,8% según datos de la CMT. Ante estas cifras es evidente que la banda ancha es una realidad cada vez más demandada y cada vez más ubicua. (Fundación Telefónica, 2011)*

La aparición de nuevos soportes de acceso a internet, como móviles y tabletas, no ha hecho más que reforzar la afirmación de *Pirolli*, que sigue vigente, y a la hora de diseñar un sistema de información deberemos tener en cuenta cuáles son las necesidades, preferencias y habilidades de nuestros usuarios. En definitiva, requerirá, por parte del diseñador, un perfil del usuario, que se dibuja a partir de la información que tenemos del mismo.

Pero, ¿de dónde obtenemos esa información?

Cuando hablamos de diseño centrado en el usuario, *las técnicas para obtener esta información son principalmente: encuestas, entrevistas y estudios de campo*. (Martín Fernández, Hassan Montero, 2004)

Entre los estudios de campo podemos incluir el análisis de la información que se genera a través de cada una de las actividades que realiza el usuario en la web.

Nuestro **objetivo** en este proyecto será analizar la actividad de recuperación de información a través de la consulta a buscadores para obtener información que nos ayude a conocer y comprender cuáles son las tendencias actuales en las búsquedas de internet, con el fin de diseñar mejores sistemas de búsqueda, implementando los motores o mejorando las interfaces de usuario, y ayudándonos también a organizar de una forma más adecuada el contenido de nuestros sistemas de información.

¿y por qué vamos a estudiar las búsquedas?

El usuario, que es el que realiza la acción de búsqueda hace un ejercicio de síntesis para describir con unas pocas palabras una necesidad de información.

En definitiva, nuestro objetivo será acercarnos al usuario recogiendo e interpretando los términos que utiliza cada vez que hace una búsqueda de contenido.

Se va a hacer uso de la *inteligencia colectiva* para obtener los datos. *La inteligencia colectiva*, como comenta Segaran, es un *término popularizado con la llegada de las nuevas tecnologías de la información* (2008, 30). Consiste, a groso modo, en recopilar datos de diferentes grupos de personas, combinarlos y analizarlos. *Aunque los métodos de inteligencia colectiva existían antes de la web, ésta, con millones de usuarios aportando información constantemente parece un buen lugar para recopilar datos* (Baeza Yates, 2011b). Baeza Yates también habla del poder de los datos en la web. Y diferencia entre datos explícitos de los que podemos obtener información semántica de forma directa,

como metadatos o etiquetas, y datos implícitos, que necesitarán ser procesados para obtener dicha información semántica.

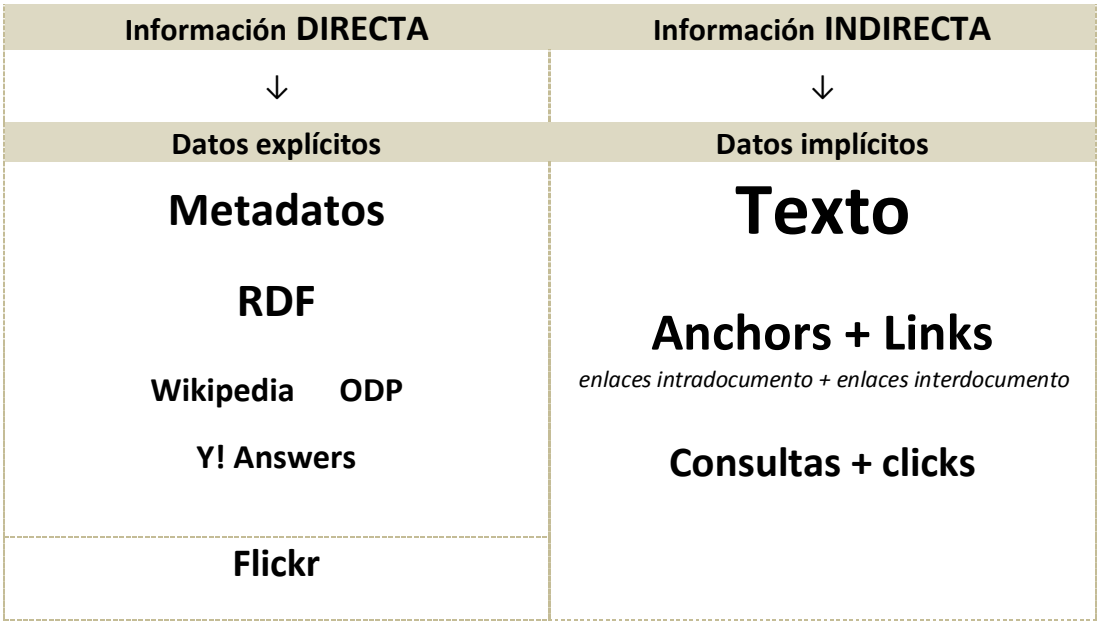


Figura 1. Información directa e indirecta. Fuente: Baeza Yates, 2011a

Baeza-Yates plantea que, por ejemplo, cada vez que buscamos algo y hacemos click en una página, podemos interpretar que estamos etiquetando la página con la consulta. Y puntualiza que en realidad estos datos no están en la web, son datos sobre lo que hacen las personas en la web. (2011b)

Y también nos habla de la calidad de los datos utilizados justificando su uso para emprender investigación basándose en los mismos: *aunque los datos generados por las personas no tienen la misma calidad que los datos editoriales, es decir los datos generados por un periódico, la cantidad de ellos es tan grande que para la misma calidad hay más datos generados por la web 2.0 que datos generados por periódicos*. (2011a)

2. JUSTIFICACIÓN DEL ESTUDIO

Desde hace más de una década se han publicado estudios para saber cómo buscan los usuarios en internet, la mayoría de las veces con la intención de implementar los motores de búsqueda y así obtener resultados que mejor satisfagan nuestra necesidad de información.

En la actualidad la tecnología al alcance del usuario es mucho más sofisticada, tenemos más recursos para participar en la red, y el uso de internet se ha incorporado de forma natural a nuestro día a día.

> *With Phase Two of the World Wide Web, we usher in a new era of online experience; the age of the amateur.[…] Web 2.0 has embraced the ordinary user with easy-to-use online tools that have enabled everyone to participate* (Kroski, 2007, 91).

El usuario de internet, haciendo uso de las herramientas de la Web 2.0 se ha convertido en catalogador, indizador y recuperador de información.

Estas son actividades propias de bibliotecarios y documentalistas, y forman parte del **proceso documental**, que es el conjunto de operaciones dirigidas a la selección, la adquisición, el registro y el tratamiento de los documentos con el fin de posibilitar su almacenamiento y recuperación para su difusión:

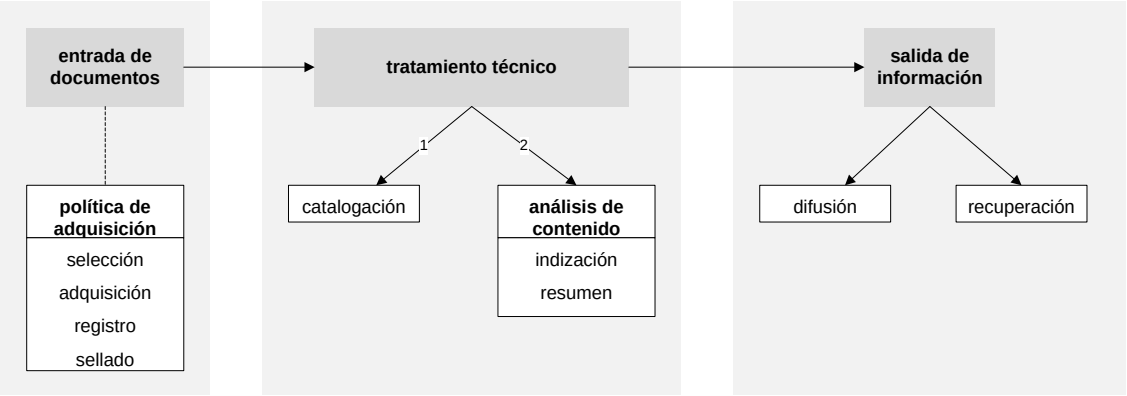


Figura 2. La cadena documental

1. La **catalogación**.

   Es el análisis de la forma, o descripción bibliográfica. El objetivo de la catalogación es *suministrar una representación del documento que lo describa de forma única, sin ambigüedades, y que permita luego identificarlo, localizarlo, incorporarlo a los ficheros*

y catálogos (Guinchat y Menou, 1983, p. 101), a partir de la información bibliográfica del propio documento.

Pero, ¿cuándo *cataloga* un usuario en internet? No es una actividad habitual, ya que es un proceso que requiere tiempo y dedicación por parte del usuario, pero en redes sociales como *Librarything* o *Entre Lectores*, en las que el objetivo es compartir información sobre nuestras bibliotecas personales, los participantes catalogan sus propios libros aportando información bibliográfica del tipo: título, autor de la obra, editorial e ISBN.

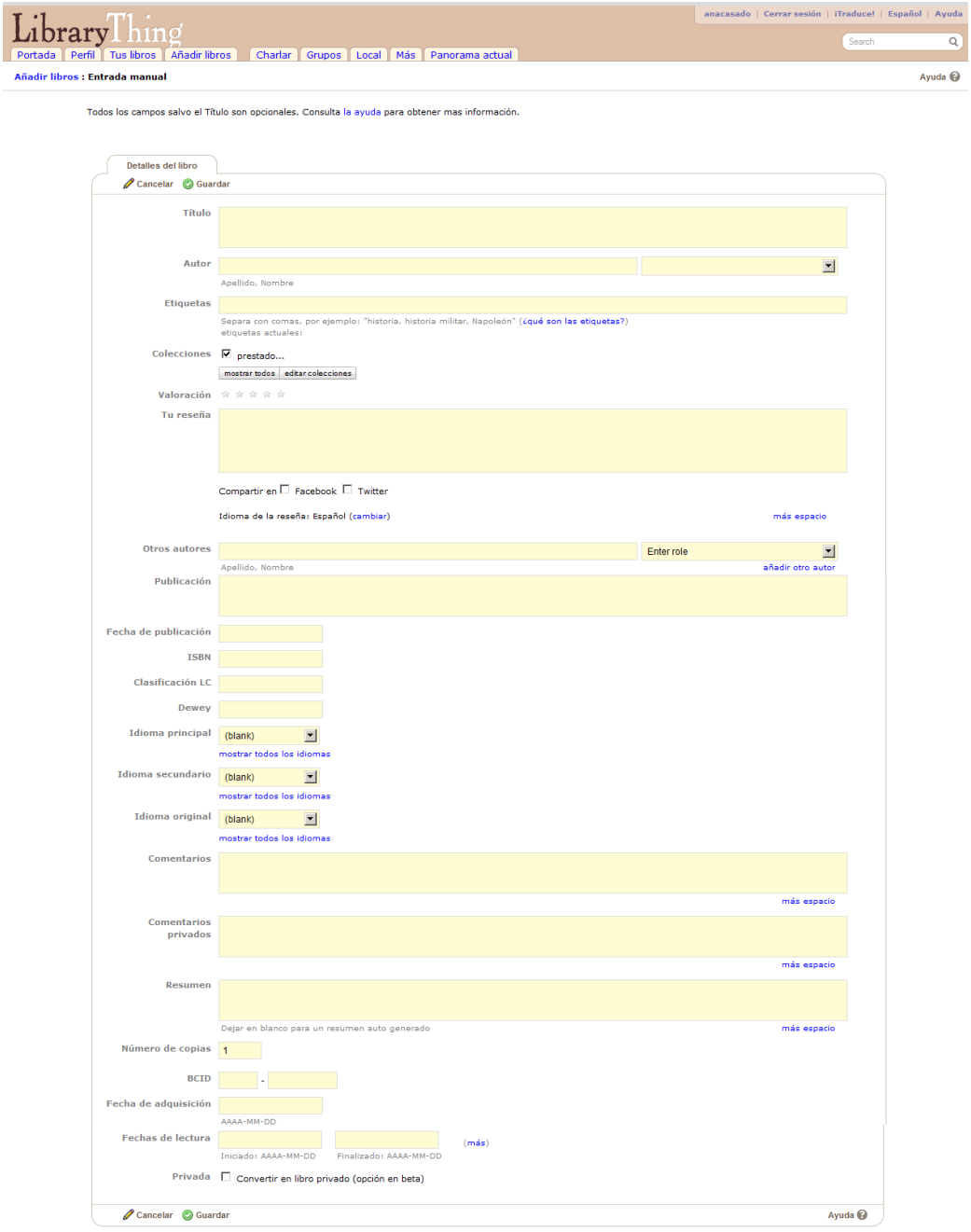


Figura 3. Interfaz de registro de libros de LibraryThing

En la Figura 3 se observa la ficha que debe rellenar el usuario al añadir un nuevo ejemplar a la colección de *Librarything*. Se cumplimentan más datos que cuando se cataloga un libro con la ISBD (reglas de catalogación). Y entre los distintos campos vemos que se entremezclan datos propios del libro, datos personales, como "valoración", "cometarios privados", "fecha de adquisición" o "fecha de lectura", y datos de descripción de contenido, como "etiquetas" y "resumen", que nos llevan a la otra actividad propia de documentalistas y bibliotecarios que ahora también realizan los usuarios de internet, la indización.

2. **Indización**

Indizar es extraer y discernir la esencia de la información y la representación del contenido a través de términos. Indizar es un proceso complejo, que consiste básicamente en examinar el documento, identificar las zonas de extracción de conceptos, en qué partes del documento están las palabras que nos darán las pistas sobre el tema, como el título, o los primeros párrafos, identificar y seleccionar los conceptos principales que representan el contenido, y después seleccionar los términos, que dependerán del tipo de indización que realicemos. La indización puede ser:

- por materias, utilizando encabezamientos de materia. Los encabezamientos de materias son un catálogo alfabético de términos que se utiliza para describir sintéticamente el asunto o tema de un documento.
- por unitérminos, utilizando sustantivos del documento, palabras que aparecen en el texto, que expresan contenido textual
- por descriptores, que son términos normalizados que expresan contenido significativo

La indización se realiza en dos momentos clave de la cadena documental, durante el tratamiento técnico documental, y en el momento de la recuperación y difusión de información. En internet, esta asignación de términos se realiza:

a. cuando el usuario describe un contenido con palabras del lenguaje natural, asignando una etiqueta o *tag* para luego poder recuperar ese contenido, poder

encontrarlo. *Los tags son términos o palabras clave en lenguaje no controlado con los que el usuario describe los recursos* (Hassan Montero, 2006a).

b. cuando busca información, a través de un buscador o a través de herramientas de navegación como las nubes de etiquetas.

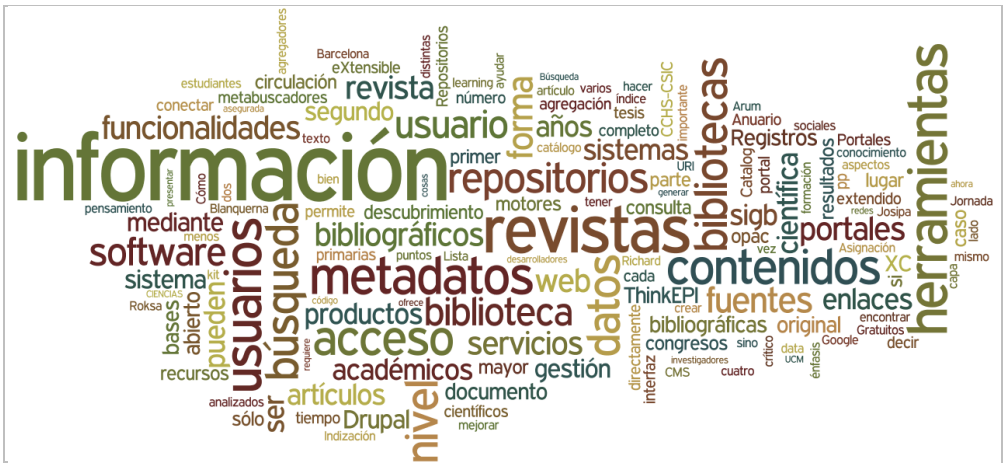


Figura 4. Nube de etiquetas realizada con la herramienta Wordle. Representa el contenido de la web http://www.thinkepi.net/

Las nubes de etiquetas son un sencillo modelo de VIRI, Interfaces gráficos para la Visualización de Información, con forma de lista ponderada de palabras clave […]. Las nubes muestran los tags o palabras clave más populares, calculándose esta popularidad por la frecuencia con la que hayan sido utilizados en la descripción de los recursos a recuperar. Una vez el usuario hace clic sobre el rótulo de un tag, obtiene una lista ordenada de recursos descritos por ese tag, así como una lista de tags relacionados con los que poder continuar su búsqueda (Hassan Montero 2006b).

3. Por eso decimos que el usuario, en internet, se convierte también en **recuperador de información**, porque durante el proceso de búsqueda, realiza una actividad de representación de información, ya sea cuando pincha sobre un término de la nube, o cuando es capaz de describir en un buscador, a través de unas pocas palabras, lo que necesita. En ese momento *el usuario, de forma activa, busca, localiza y accede a la información que necesita, ya que tiene una representación sintáctica de sus necesidades de información* (Hassan Montero, 2006a)

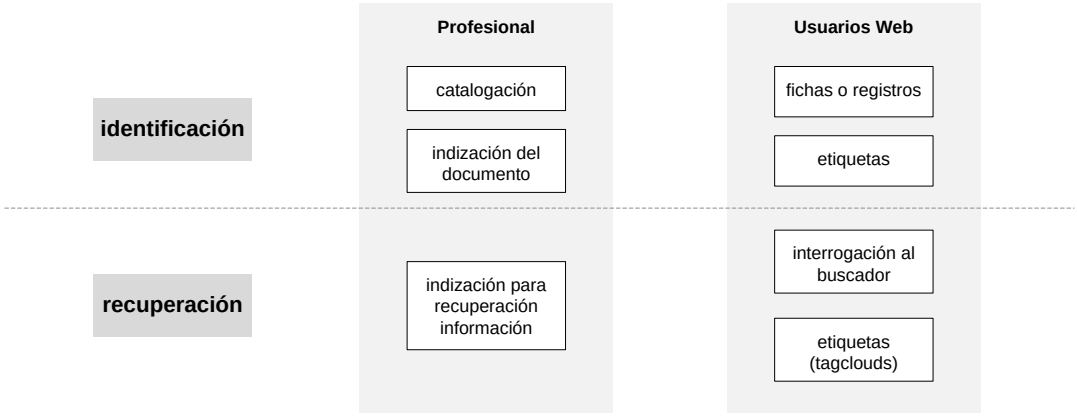


Figura 5. Comparación actividades de profesional información vs. usuario

Extraer datos sobre búsquedas y etiquetado, conocer los criterios que sigue el usuario para realizar cada una de estas actividades, nos dará información con la que podremos mejorar los servicios de recuperación de información diseñados para él (tanto en bibliotecas como en servicios de información de otras instituciones). Y también nos podrá dar argumentos a la hora de tomar decisiones sobre organización de contenidos, herramientas de navegación, terminología, diseño de interfaces web,… todo ello dirigido a que el usuario encuentre y acceda a la información que necesita en cada momento.

Teniendo en cuenta el volumen de información que podemos encontrar en la web y los datos del artículo de Jansen y Booth que demuestran que dependiendo del tema, la intención de búsqueda del usuario es diferente (2010), creemos necesario acotar nuestro campo de estudio.

En este estudio el entorno elegido ha sido el bibliotecario:

1º. Por su relación con los estudios de Documentación en los que se enmarca este Proyecto Final de Carrera. La Licenciatura en Documentación es el segundo ciclo natural de los estudios de Biblioteconomía.

2º. Porque en el nuevo papel que juega el usuario en internet, realiza tareas propias del profesional de la información al utilizar las herramientas que la web 2.0 pone a su disposición para participar en la web.

   La información que extraigamos de las búsquedas y el etiquetado que realiza el usuario puede ser estudiada en relación con los vocabularios controlados que utiliza el profesional de la información para describir contenido.

3º. Porque cualquier persona es un usuario potencial de la biblioteca pública, que no responde a un perfil de usuario con ningún nivel de conocimientos determinado. Todos podemos acceder a los servicios de la biblioteca pública, igual que cualquiera puede ser usuario de internet.

   Según la definición elaborada por IFLA/UNESCO, "biblioteca pública es una organización establecida, apoyada y financiada por la comunidad, tanto a través de una autoridad u órgano local, regional o nacional o mediante cualquier otra forma de organización colectiva. Proporciona acceso al conocimiento, la información y las obras de creación gracias a una serie de recursos y servicios y está a disposición de todos los miembros de la comunidad por igual, sean cuales fueren su raza, nacionalidad, edad, sexo, religión, idioma, discapacidad, condición económica, laboral y nivel de instrucción." (Ministerio de Cultura, 2011)

4º. Porque existe una necesidad real de conocer los intereses de los usuarios potenciales de las bibliotecas. Con el protagonismo que ha adquirido el usuario con la web 2.0, y los cambios tecnológicos que han cambiado radicalmente nuestra forma de acceso y uso de la información, las bibliotecas han iniciado un camino de tecnificación que supone transformar los servicios presenciales tradicionales en nuevos servicios a través de medios electrónicos.

   Antes la biblioteca seleccionaba libros, ofertaba servicios, información y orientación al usuario y mantenía intacto sus monopolios. Ahora tiene que compartir protagonismo y desarrollar nuevas funciones acorde con los tiempos que corren (Lozano, 2010)

   Identificar las demandas de su usuario será fundamental para ofrecer unos servicios acordes a sus necesidades.

5º. Se plantea la necesidad de establecer un contexto para las búsquedas, para poder así entenderlas y extraer información de las mismas.

   There is not information without context (Baeza-Yates, 2008).

Y los datos que vamos a analizar son las búsquedas que los usuarios realizan en Google relacionadas con este entorno.

Antes de continuar ¿a qué datos nos referimos cuando hablamos de palabras clave y búsquedas?

- **Palabra clave (*keywords*)**: palabra o frase que los usuarios introducen en un buscador para satisfacer una necesidad de información. La expresión que utiliza un usuario para interrogar al buscador.
- **Búsqueda (*queries*)**: es la acción que interrogar al buscador utilizando una palabra clave determinada.

Una vez elegido el entorno en el que vamos a trabajar nos surgen otras preguntas. ¿Qué información nos ofrecen las búsquedas? ¿En qué medida nos ayudan a "conocer mejor al usuario" y cuáles son esos puntos en común?

No se trata de estudiar si el usuario conoce el nombre de un autor, o si está al día de las últimas novedades editoriales. Esa información podrá encontrarla en el sistema de información adecuado, pero sí podemos averiguar si al realizar búsquedas utiliza nombres propios, o si busca por temas; y si busca por temas, si existe algún paralelismo entre éstos y los encabezamientos de materia que utilizan las bibliotecas públicas para describir el contenido de un documento. Las búsquedas y las etiquetas también nos pueden dar información sobre formatos y tipo de documento que busca el usuario. Puede ser un dato relevante el porcentaje de búsquedas que hacen referencia a soportes. Nos puede ayudar a tomar decisiones sobre cómo organizar el contenido en nuestro sistema de información, por temas, por formato, por novedad,… y podemos averiguar si existe una tendencia a utilizar formatos electrónicos, por ejemplo, o si el formato CD-ROM ha quedado ya obsoleto,…

Y la información que obtengamos de las búsquedas nos puede ayudar también a mejorar nuestras propias herramientas: podemos actualizar nuestra lista de encabezamientos de materia o implementar las herramientas de recuperación de información con los términos que utiliza el usuario.

Figura 6. Página de resultados de búsqueda del catálogo de la UCA

Hay bibliotecas, como la de la Universidad de Cádiz, que han implementado sus sistemas de recuperación de información con herramientas de la web 2.0, como nubes de etiquetas que se nutren de los encabezamientos de materia y de las etiquetas que asignan los usuarios a los documentos (http://www.uca.es/area/biblioteca).

Algunas etiquetas son muy genéricas y pueden aplicarse a muchos documentos o a distintos contenidos web, pero la combinación de estas etiquetas con otras más específicas nos permitirá describir y recuperar contenido específico. Aún así, *debido a la naturaleza más genérica que específica de los tags*, *la indización social sería más apropiada para ofrecer sistemas de navegación* (Hassan Montero, 2006a). Si somos capaces de incorporar esta información, a nuestro sistema de información podremos tener una herramienta de navegación y desambiguación de búsquedas, como el catálogo de la UCA, y también nos aportará información acerca del nivel que tienen nuestros usuarios sobre temas específicos o contenido que podamos ofrecer a través de nuestro sistema de información.

Otra información que podemos obtener al estudiar las búsquedas, es la intención con la que se realizan. Cuando Jansen habla de intención de búsqueda se *refiere a la forma en la*

que el usuario expresa el objetivo de su búsqueda (2010). Por la expresión que utilice podremos entender que busca un tipo u otro de información.

La intención de búsqueda se puede clasificar, siguiendo la propuesta de Broder (2002) en:

a. Navegacional: Aquellas búsquedas que explícitamente hacen referencia a una institución, o a una información concreta, albergada en páginas concretas. El usuario busca llegar a algún sitio en particular.
b. Transaccional: El objetivo del usuario, no es encontrar o localizar contenidos informativos, sino realizar alguna actividad en la web, transacciones o acciones de intercambio.
c. Informativa: La intención es adquirir alguna información que se supone puede estar presente en una o más páginas web.

   En el caso de las búsquedas informativas, la precisión con la que formulemos nuestra petición, y lo específicos que hayamos sido en el etiquetado, nos dará unos resultados u otros. Siempre teniendo en cuenta que no siempre está disponible el documento que buscamos, y que en muchas ocasiones no buscamos una información concreta.

Conocer la intención de búsqueda *se puede aprovechar para clasificar mejor la intención del usuario de base y mejorar el rendimiento de los motores de búsqueda Web.* (Jansen, 2010). Y también para tomar decisiones a la hora de determinar los componentes de un sistema de información, que podemos dividir en:

a. sistemas organizativos: cómo categorizar la información, por ejemplo, por materias, cronológicamente,…

b. sistemas de etiquetado: cómo representamos la información, por ejemplo, con terminología científica o común.

c. sistemas de navegación: cómo navegamos o nos movemos a través de la información, por ejemplo, recorriendo una jerarquía.

d. Sistemas de búsqueda: cómo buscamos información, por ejemplo, realizando una consulta sobre un índice.

(Morville, Rosenfeld, 2006, 49)

En cada caso deberemos conocer cuál es el nivel de conocimientos de nuestro usuario potencial respecto a la información que ofrece nuestra web o sistema de información, qué tipo de información es la más demanda, la que tiene más interés, cómo realiza las búsquedas, la experiencia de nuestros usuarios, y nos podremos apoyar en la información

que tengamos al respecto para tomar decisiones sobre la arquitectura de información de la web que estemos diseñando.

2.1. La jerarquía de la información

Cuando un usuario accede a una página web, lo primero que hace es echar un vistazo general, una exploración visual, en la que se detectan cuáles son los contenidos de información más importantes, cómo se ha estructurado la información, por los menús, submenús, cabeceras, que hacen que detectemos cierta ordenación jerárquica del contenido de la página.

Si echamos un vistazo a las páginas de inicio de las redes de bibliotecas de algunas comunidades autónomas, Madrid, Cataluña, Comunidad Valenciana y Navarra, por ejemplo, obtendremos resultados dispares. Algunos resultados nos llevan al catálogo en línea de la comunidad autónoma, otros a la página de inicio de las bibliotecas de la comunidad, otras solo nos dan direcciones. De entrada vemos que hay múltiples respuestas para una misma pregunta según la comunidad, que tiene que ver con el lugar que ocupan las bibliotecas públicas en la jerarquía administrativa en cada comunidad. Una vez seleccionadas lo que entendemos que son las páginas de inicio de las Bibliotecas Públicas de cada comunidad, podemos observar que se han tomado decisiones distintas a la hora de plantear el diseño de las webs:

Figura 7. Página de inicio de las Bibliotecas públicas de la Comunidad de Madrid

Figura 8. Página de inicio de las Bibliotecas públicas de Cataluña

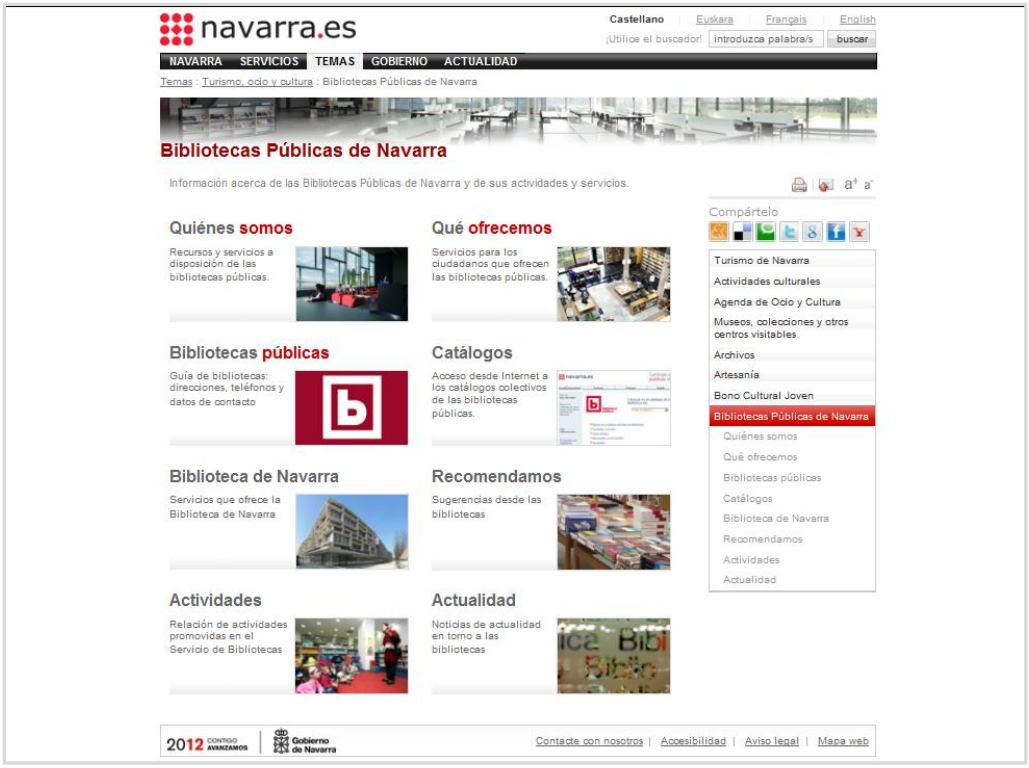


Figura 9. Página de inicio de la Red Bibliotecas públicas de Navarra

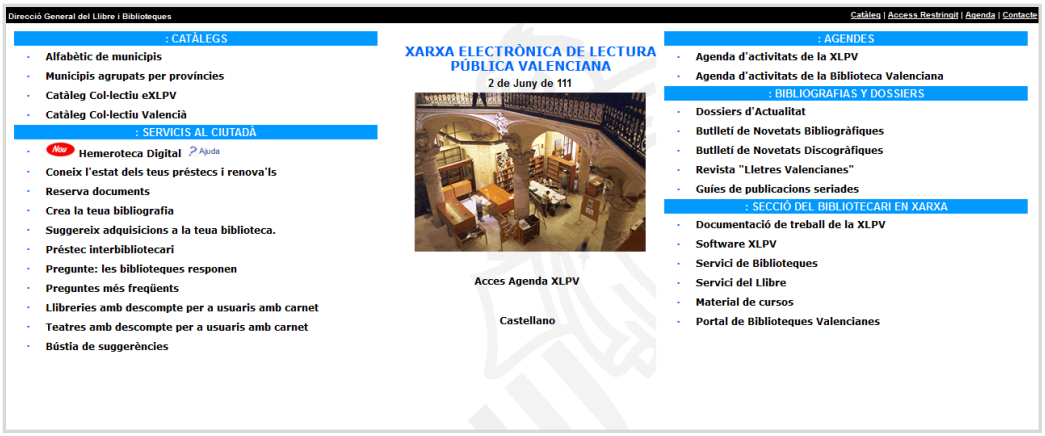


Figura 10. Página de inicio de la Red electrónica de lectura pública valenciana

No vamos a entrar aquí a analizar cada una de las webs, pero a simple vista observamos diferencias notables ya en las páginas de inicio de cada una de las cuatro que aquí mostramos, en el volumen de información ofrecido en cada una de ellas, el modo de estructurar los servicios, la importancia que se le da al catálogo, y también en los servicios que a simple vista ofrece cada una de ellas…

Un mayor conocimiento de cómo buscan los usuarios y cómo etiquetan, nos pemitiría generar una mejor jerarquía de información. Si encontramos una tendencia clara en cómo buscan y etiquetan los usuarios, no debería haber portales tan dispares para temáticas y tareas tan concretas como las de portales tan similares, que en teoría se dirigen a un público similar, y ofrecen los mismos servicios básicos.

Morville y Rosendelf definían Arquitectura Web (1998) como:

- La combinación de la organización, etiquetado y los esquemas de navegación dentro de un sistema de información
- El diseño estructural de un espacio de información para facilitar las tareas de acabado y acceso intuitivo a los contenidos.
- El arte y ciencia de estructurar y clasificar sitios web e intranets para ayudar a los usuarios a encontrar y administrar su información.
- Una disciplina emergente y una comunidad práctica enfocada en traer los principios de diseño y arquitectura a los entornos digitales.

Nos hablan de entornos de información compartida, de distintos aspectos que se combinan en el sitio web, el tipo de organización, los sistemas de navegación, las búsquedas,… las decisiones que tomemos tendrán que tener en cuenta cada uno de

estos aspectos, y las tendremos que tomar apoyándonos en la información que tengamos sobre nuestra organización y sobre nuestros usuarios.

Ronda León va un paso más allá en la definición y apunta que e*l nuevo enfoque de la arquitectura de información radica en la atención que le presta a la gestión de la información (organizar, estructurar, representar) y a las necesidades de usuarios* (2005) estableciendo una relación directa entre las disciplinas de ciencias de la documentación y la arquitectura de la información que postula como una nueva disciplina dentro de la primera.

3. ESTADO DE LA CUESTIÓN

Hasta el momento hemos tratado superficialmente varios temas relacionados con la descripción y recuperación de la información: búsquedas en la web, etiquetado social, intención de búsqueda, descripción de contenidos, etc. Llegados a este punto se hace necesario profundizar sobre los aspectos concretos en los que se centra este estudio: las operaciones de búsqueda de información por parte del usuario.

Primero analizaremos la bibliografía de los artículos con los que nos hemos documentado para plantear este estudio en forma de línea de tiempo y acoplamiento bibliográfico. Esto nos dará una perspectiva general del camino recorrido en cada tema y de las líneas de estudio en las que nos hemos centrado y en las que se enmarca nuestro estudio.

Y en segundo lugar se hace un repaso de los artículos publicados sobre búsquedas web, artículos que han sido nombrados en diferentes revisiones publicadas sobre cada tema.

3.1. Recopilación de información. Metodología

Se realizan las primeras búsquedas en Google Académico, http://scholar.google.es/, utilizando términos genéricos, sobre todo para localizar autores, y hacer una evaluación rápida del volumen de publicaciones sobre el tema.

Términos utilizados para hacer estas primeras búsquedas: *searching and tagging*, *web searching, web searching data, searching the web, tagging, social tagging*

Los primeros artículos con los que trabajamos son:

- *Strong Regularities in World Wide Web Surfing*, de Bernardo A. Huberman, Peter L. T. Pirolli, James E. Pitkow and Rajan M. Lukose, que se centra en la navegación
- *How are we searching the World Wide Web? A comparison of nine search engine transaction logs*, de Bernard J. Jansen y Amanda Spink, que analiza los términos que utilizamos para interrogar al buscador.

Una vez localizados los primeros artículos, y a través de sus palabras clave, se realizaron nuevas búsquedas: *browsing retrieval, search intention, searching data semantic web, definición de servicio web,…*

Se utilizó la bibliografía para localizar más artículos de interés en el tema que estamos estudiando.

También utilizamos los términos de búsqueda en Google, y localizamos algunas webs y blogs relacionados con el tema de las que también hemos obtenido información pertinente:

- No solo usabilidad, http://www.nosolousabilidad.com/
- The long Tail, http://www.longtail.com/the_long_tail/
- Grupo ThinkEpi, http://www.thinkepi.net/
- The MIT Media Lab, http://www.media.mit.edu/
- Wikipedia, http://es.wikipedia.org/wiki/Web_2.0
- Searchblog, http://battellemedia.com/
- Maestros del Web, http://www.maestrosdelweb.com/editorial/web2/

Acceso a las publicaciones científicas:

Algunos artículos se pueden descargar directamente en Google Académico, son de libre acceso. Pero muchos de los artículos localizados son de acceso restringido, y la mayoría están publicados en la Biblioteca Digital de ACM, http://dl.acm.org/, por lo que se han podido descargar a través del catálogo de la biblioteca de la UPV, desde los ordenadores del campus.

Otra fuente de información a la que hemos llegado a través de las búsquedas ha sido DBLP Computer Science Bibliography, http://www.informatik.uni-trier.de/~ley/db/, que *es un servidor de información bibliográfica sobre las principales revistas de ciencias de la* computación *y otros procedimientos. DBLP indexa más de un millón de artículos y contiene más de 10.000 enlaces a páginas de inicio de científicos de la computación. DBLP se produce en el departamento de Ciencias de la Computación de la Universidad de Trier, en Alemania.*

De este servidor hemos obtenido principalmente información relativa a coautorías.

3.2. Análisis de la bibliografía

A partir de la bibliografía recogida en los artículos citados en este trabajo hemos realizado una línea de tiempo que representa las publicaciones relacionadas con la búsqueda y recuperación de información, que tiene un mayor auge en los años 90, década que podríamos identificar como el origen histórico de la *Era de la Documentación Digital* (Rovira, 2004, 9) y con las publicaciones relacionadas con el etiquetado social.

3.2.1. Búsquedas. Línea de tiempo

Línea de tiempo dibujada con la bibliografía de los artículos que hemos seleccionado para estudiar cómo buscamos y la intención de búsqueda de los usuarios.

Interpretación de la línea de tiempo:

- En la bibliografía hay dos referencias que no se han representado por una cuestión de espacio, ya que datan de 1945 y 1949.
- De los títulos de *los* artículos de la década de 1970, podemos intuir que tratan de establecer las bases teóricas para conseguir eficiencia de los sistemas de recuperación de información: *relevance, retrieval characteristics, relevance weighting, co-occurrence data, probabilistic models, theory of indexing, optimum term weighting,…*
- De los 80 encontramos pocas referencias bibliográficas. Y las que tenemos son sobre artículos *genéricos* o estudios experimentales sobre el uso de sistemas de información.
- En 1989 aparece la primera referencia a un estudio del *log* de un OPAC.
- Los años en blanco o con una producción significativamente más baja no coinciden necesariamente con que no haya habido producción. Es el caso de 1994 o 2009. Es difícil pensar que esos años no se publicara ningún artículo relacionado con el tema.
- Pero sí parece significativo el volumen de artículos referenciados que se publicaron a finales de los años 90 y primeros 2000. Son artículos que analizaban las consecuencias de la aparición de los primeros buscadores.

El primer buscador fue Wandex, […] un robot desarrollado por Mattew Gray en el MIT, en 1993. Otro de los primeros buscadores, Aliweb, también apareció en 1993 y todavía está en funcionamiento. El primer motor de búsqueda de texto completo fue WebCrawler, que apareció en 1994. A diferencia de sus predecesores, éste permitía a sus usuarios una búsqueda por palabras en cualquier página web, lo que llegó a ser un estándar para la gran mayoría de los buscadores. También apareció en 1994 Lycos.[…]

Muy pronto aparecieron muchos más buscadores, como Excite, Infoseek, Inktomi, Northern Light y Altavista. De algún modo, competían con directorios (o índices temáticos) populares tales como Yahoo!. Más tarde, los directorios se integraron o se añadieron a la tecnología de los buscadores para aumentar su funcionalidad. (Colaboradores de Wikipedia). Y por fin, en 1998, apareció Google.

- Después del *boom de publicaciones* observamos que sigue habiendo publicaciones relacionadas con el tema. Se mantiene una publicación constante en la década del 2.000, y los artículos sirven de referencia para publicaciones posteriores.

LÍNEA DE TIEMPO. ESTUDIOS DE BÚSQUEDA DE INFORMACIÓN

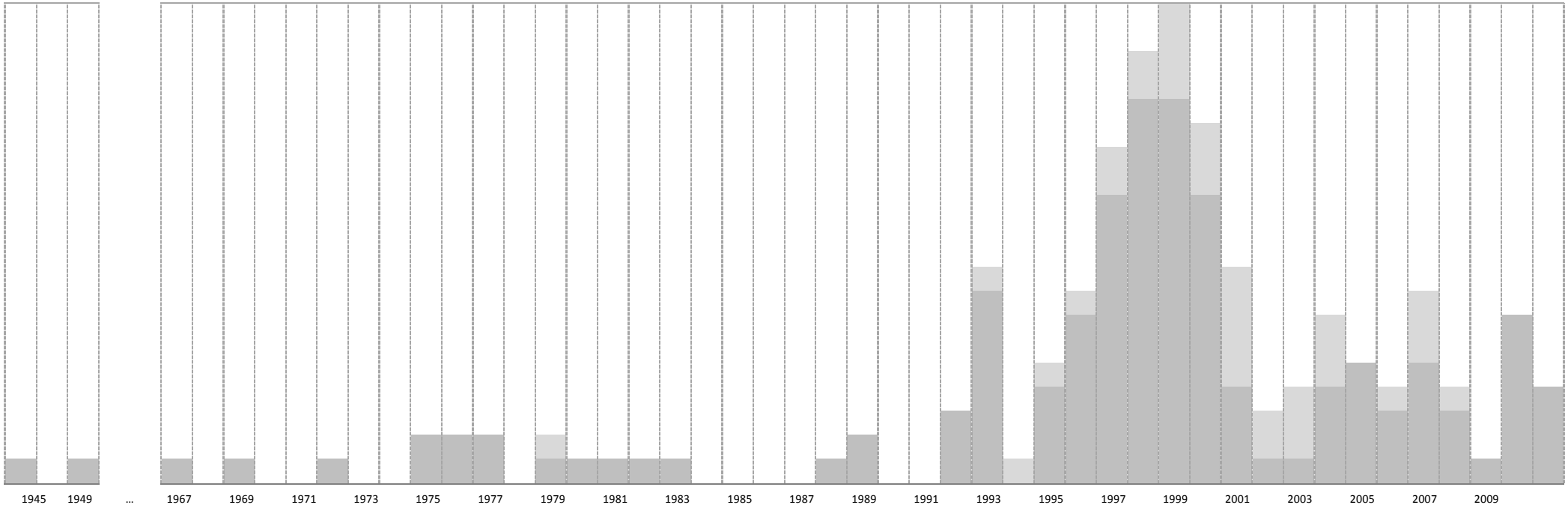


Figura 11. Línea de tiempo. Representación gráfica de estudios sobre búsqueda de información

Leyenda

Artículos sobre búsquedas

Artículos sobre intención de búsqueda

Bibliografía utilizada para realizar la gráfica en Anexo 1

3.2.2. Búsquedas. Acoplamiento Bibliográfico

El acoplamiento bibliográfico se produce cuando 2 documentos comparten 1 ó más referencias bibliográficas. El acoplamiento nos habla sobre el documento de referencia, y sirve para detectar líneas de investigación comunes.

Estas son las referencias compartidas por 2 o más documentos de la bibliografía de referencia relacionada con las búsquedas:

1. Gordon, M. and Pathak, P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing and Management, 35* (2) 141-180.

   Referencia bibliográfica en:

   - a. Searching the web: the public and their queries
   - b. A Review of Web Searching Studies and a Framework for Future Research

2. Huberman, B.A., Pirolli, P., Pitnow, J.E. and Lukose, R.M. (1998) Strong regularities in World Wide Web surfing. Science, 280 (5360) 95-97.

   Referencia bibliográfica en:

   - a. Searching the web: the public and their queries
   - b. A Review of Web Searching Studies and a Framework for Future Research
   - c. Real life, real users, and real needs: a study and analysis of user queries on the web
   - d. Real life information retrieval: a study of user queries on the web

3. Jansen, B.J., Spink, A., Bateman, J. and Saracevic, T. (1998). Real life information retrieval: A study of user queries on the Web. SIGIR Forum, 33 (1) 5-17.

   Referencia bibliográfica en:

   - a. Searching the web: the public and their queries
   - c. Real life, real users, and real needs: a study and analysis of user queries on the web
   - f. Análisis de las búsquedas realizadas, categorías accedidas y documentos vistos en un directorio Web

4. Jones, S., Cunningham, S.J., and McNab, R. Usage analysis of a digital library. Proceedings of the Third ACM Conference on Digital Libraries (June 1998) 293-294.

   Referencia bibliográfica en:

   - a. Searching the web: the public and their queries
   - c. Real life, real users, and real needs: a study and analysis of user queries on the web

5. Lawrence, S. and Giles, C.L. Accessibility of information on the web. Nature, 400, (July 8, 1999) 107-109.

   Referencia bibliográfica en:

   - a. Searching the web: the public and their queries
   - b. A Review of Web Searching Studies and a Framework for Future Research
   - e. How are we searching the world wide web? a comparison of nine search engine transaction logs

6. Ross, N. C. M., & Wolfram, D. (in press). End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine. Journal of the American Society for Information Science.

   Referencia bibliográfica en:

   - a. Searching the web: the public and their queries
   - b. A Review of Web Searching Studies and a Framework for Future Research
   - e. How are we searching the world wide web? a comparison of nine search engine transaction logs

7. Zipf, G.K. (1949). Human Behavior and the Principle of Least Effort. Cambridge: Addison-Wesley.

   Referencia bibliográfica en:

   - a. Searching the web: the public and their queries
   - g. Strong regularities in world wide web surfing

8. Nielsen Media. (1997). Search engines most popular method of surfing the web [Website*]. Commerce Net/Nielsen Media.*

   Referencia bibliográfica en:

   - b. A Review of Web Searching Studies and a Framework for Future Research

- e. How are we searching the world wide web? a comparison of nine search engine transaction logs

9. Croft, W. B., Cook, R., & Wilder, D. (1995). *Providing government information on the Internet: experiences with THOMAS. In Proceedings of Digital Libraries '95 Conference, Austin, TX* (pp. 19-24).

   Referencia bibliográfica en:

   - b. A Review of Web Searching Studies and a Framework for Future Research
   - c. Real life, real users, and real needs: a study and analysis of user queries on the web

10. Crovella, M., & Bestavros, A. (1996). Self-similarity in World Wide Web traffic evidence and possible causes. Paper presented at ACM SIGMETRICS Conference on Measurement & Modeling of Computer Systems, Philadelphia, PA

    Referencia bibliográfica en:

    - b. A Review of Web Searching Studies and a Framework for Future Research
    - c. Real life, real users, and real needs: a study and analysis of user queries on the web
    - d. Real life information retrieval: a study of user queries on the web

11. Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. Journal of the American Society for Information Science, 44(3), 161–174

    Referencia bibliográfica en:

    - b. A Review of Web Searching Studies and a Framework for Future Research
    - c. Real life, real users, and real needs: a study and analysis of user queries on the web
    - d. Real life information retrieval: a study of user queries on the web

12. Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. Information Processing & Management, 36(2), 207–227.

    Referencia bibliográfica en:

    - b. A Review of Web Searching Studies and a Framework for Future Research

- e. How are we searching the world wide web? a comparison of nine search engine transaction logs

13. Kehoe C., Pitkow J., Morton K. 1997. GVU's 8th WWW user survey. Atlanta, GA: Graphic, Visualization, and Usability Center, Georgia Tech Research Center.

    Referencia bibliográfica en:

    - b. A Review of Web Searching Studies and a Framework for Future Research
    - c. Real life, real users, and real needs: a study and analysis of user queries on the web
    - d. Real life information retrieval: a study of user queries on the web

14. Lawrence, S., & Giles, C. L. (1998). Searching the World Wide Web. Science, 280(5360), 98-100.

    Referencia bibliográfica en:

    - b. A Review of Web Searching Studies and a Framework for Future Research
    - c. Real life, real users, and real needs: a study and analysis of user queries on the web
    - d. Real life information retrieval: a study of user queries on the web

15. Lynch, C. (1997). Searching the Internet. Scientific American, 276(3), 52–56

    Referencia bibliográfica en:

    - b. A Review of Web Searching Studies and a Framework for Future Research
    - c. Real life, real users, and real needs: a study and analysis of user queries on the web
    - d. Real life information retrieval: a study of user queries on the web

16. Peters, T. A. (1993). The history and development of transaction log analysis. *Library Hi Tech, 42*(11:2), 41-66.

    Referencia bibliográfica en:

    - b. A Review of Web Searching Studies and a Framework for Future Research
    - c. Real life, real users, and real needs: a study and analysis of user queries on the web
    - e. How are we searching the world wide web? a comparison of nine search engine transaction logs

17. Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large web search engine query log. SIGIR Forum, 33(1), 6–12.

Referencia bibliográfica en:

- b. A Review of Web Searching Studies and a Framework for Future Research
- e. How are we searching the world wide web? a comparison of nine search engine transaction logs

18. Spink, A., Greisdorf, H., & Bateman, J. (1998). From highly relevant to not relevant: examining the different regions of relevance. *Information Processing and Management, 34*(5), 599-622.

    Referencia bibliográfica en:

    - b. A Review of Web Searching Studies and a Framework for Future Research
    - c. Real life, real users, and real needs: a study and analysis of user queries on the web

19. Jansen, B. J., & Spink, A. (Forthcoming). An analysis of web searching by European alltheweb.Com users. *Information Processing & Management.*

    Referencia bibliográfica en:

    - e. How are we searching the world wide web? a comparison of nine search engine transaction logs
    - h. The Intention Behind Web Queries

20. Jansen, B. J., Spink, A., Bateman, J., & Saracevic, T. (1998a). Searchers, the Subjects They Search, and sufficiency: A Study of a Large Sample of Excite Searches. In Proceedings of WebNet 98 Conference, Orlando, FL, November 1999.

    Referencia bibliográfica en:

    - c. Real life, real users, and real needs: a study and analysis of user queries on the web
    - f. Análisis de las búsquedas realizadas, categorías accedidas y documentos vistos en un directorio Web

21. Baeza-Yates, R.; Ribeiro-Neto, B. (1999). Modern Information Retrieval. Addison-Wesley, Wokingham, UK, 1999.

    Referencia bibliográfica en:

    - f. Análisis de las búsquedas realizadas, categorías accedidas y documentos vistos en un directorio Web

- i. Indización social y recuperación de información

22. Abdulla, G., Fox E.A., & Abrams, M. (1997). Shared User Behavior on the World Wide Web. *Proceedings of the WebNet'97*, 54-59.

    Referencia bibliográfica en:

    - c. Real life, real users, and real needs: a study and analysis of user queries on the web
    - d. Real life information retrieval: a study of user queries on the web

23. Bates, M.J., Wilde, D. N. and Siegfried, S. (1993) An analysis of search terminology used by humanities scholars: *The Getty online searching project report. Library Quarterly, 63* (1), 1-39.

    Referencia bibliográfica en:

    - c. Real life, real users, and real needs: a study and analysis of user queries on the web
    - d. Real life information retrieval: a study of user queries on the web

24. Fenichel, C. H. (1981). Online searching: Measures that discriminate among users with different.types of experience. *Journal of the American Society for Information Science,* 32, 23-32.

    Referencia bibliográfica en:

    - c. Real life, real users, and real needs: a study and analysis of user queries on the web
    - d. Real life information retrieval: a study of user queries on the web

25. Fenichel, C. H. (1981). Online searching: Measures that discriminate among users with different.types of experience. *Journal of the American Society for Information Science,* 32, 23-32.

    Referencia bibliográfica en:

    - c. Real life, real users, and real needs: a study and analysis of user queries on the web
    - d. Real life information retrieval: a study of user queries on the web

26. Saracevic, T. (1997). Users lost: Reflections on the past, future, and limits of information science*. SIGIR Forum, 31* (2) 16-27

Referencia bibliográfica en:

- c. Real life, real users, and real needs: a study and analysis of user queries on the web
- d. Real life information retrieval: a study of user queries on the web

27. Spink, A. & Saracevic, T. (1997). Interactive information retrieval: Sources and effectiveness of search terms during mediated online searching. Journal of the American Society for Information Science, 48, (8), 741-761.

    Referencia bibliográfica en:

    - c. Real life, real users, and real needs: a study and analysis of user queries on the web
    - d. Real life information retrieval: a study of user queries on the web

28. Mathes, A. (2004). Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Unpublished paper.

    Referencia bibliográfica en:

    - i. Indización social y recuperación de información
    - j. Structure and form of folksonomy tags: the road to the public library catalogue

29. Quintarelli, E. (2005). Folksonomies: power to the people. ISKO Italy-UniMIB meeting: Milan. June 24, 2005.

    Referencia bibliográfica en:

    - i. Indización social y recuperación de información
    - j. Structure and form of folksonomy tags: the road to the public library catalogue

El artículo con mayor acoplamiento bibliográfico es el 2, *Strong regularities in World Wide Web surfing*, citado por 4 artículos de nuestra bibliografía (a, b, c, d).

Le siguen el **3** y el **16**, que han sido citados por (a, c, f) y (c, b, e) respectivamente.

Luego tenemos grupos de artículos que son citados por los mismos artículos de la bibliografía.

Es el caso de los artículos [10, 11, 13, 14, 15], que han sido citados por (b, c, d), y por ningún otro artículo. Observamos que los artículos citados son de diferentes autores, pero con puntos en común: todos son artículos publicados en la década de los 90, y no necesariamente tratan sobre búsquedas en internet. Más bien la línea común de estos artículos es la observación de las conductas de los usuarios en el nuevo medio y el estudio de las posibilidades que ofrece para establecer un contacto directo con el usuario. El artículo de Kehoe y Pitkow analiza 5 encuestas en línea, la primera de 1994, proponiendo una metodología para extraer datos de calidad que puedan ser útiles en estudios posteriores.

Y los tres artículos que citan a este grupo tienen en común al autor, Bernard J. Jansen, con lo que podemos extraer que los tres pertenecen a una misma línea de investigación del autor, las búsquedas en la web.

Lo mismo ocurre con los artículos [5, 6], los 2 han sido citados por (a, b, e). La temática de nuestra bibliografía es clara, las búsquedas en la web. Los artículos citados se interesan por la accesibilidad a la información y la coocurrencia de pares de términos en las búsquedas de temas específicos en internet.

Los dos artículos han sido citados en publicaciones de Jansen, que sigue una línea de estudio clara, los usuarios y las búsquedas en la web.

3.3. Estudios sobre el tema

3.3.1. Estudios sobre análisis de búsquedas

Desde finales de los años 60 se publican estudios que evalúan la eficacia de los sistemas de información y proponen varemos para evaluar el rendimiento de los mismos, se desarrollan teorías y se proponen nuevos modelos para diseñar sistemas de información basándose en la exhaustividad y la precisión de los resultados de búsqueda, y midiendo la relevancia de los términos que se utilizan tanto en la descripción como en la recuperación de información para obtener resultados pertinentes.

En los años 80 entran en las bibliotecas y centros de documentación los primeros microordenadores de razonable potencia y los primeros sistemas de reconocimiento OCR (Rovira et al., 2004, 9). En esta década aparecen los primeros estudios de archivos log de un OPAC.

Las publicaciones relacionadas con la recuperación de información han estado condicionadas por un entorno cambiante, cambiante por la evolución de la tecnología que ha supuesto, a partir de los años 90, una transformación definitiva del consumo y del acceso a la información por parte del usuario, que asume un papel activo en el proceso de recuperación de información y condiciona el diseño de los sistemas de información, que a la hora de ser concebidos tendrán que valorar tanto las nuevas capacidades del usuario como las herramientas que sobre todo la web 2.0 ha puesto a su alcance.

Jim Jansen, nos ayuda a entender este entorno cambiante, a través de algunas de las personalidades más relevantes para él en el desarrollo de las búsquedas patrocinadas y que más han influido en los cambios que se han sucedido en torno a la búsqueda de información (Jansen, 2011):

- *Scott Banister, Jeffrey Brewer, and Bill Gross – credited with conceiving the idea of keyword advertising*

  Scott Banister, cofundador en 1995 de *SponsorNet New Media, Inc.* y creador en 1996 de *Submit It!*, cofundador en 2000 de *IronPort* y cofundador, asesor y miembro de la junta en *PayPal* entre otras; Jeffrey Brewer, fundador de Goto.com y Bill Gross, fundador de *Idealab*, desarrollan el concepto de *Pay per CLick*: entre los tres conciben la idea de **publicidad por palabras clave**.

- *John Battelle – the Database of Intentions*

  publicado el 13 de noviembre de 2003

  The Database of Intentions is simply this: The aggregate results of every search ever entered, every result list ever tendered, and every path taken as a result. It lives in many places, but three or four places in particular hold a massive amount of this data (ie MSN, Google, and Yahoo). This information represents, in aggregate form, a place holder for the ***intentions of humankind*** *- a massive database of desires, needs, wants, and likes that can be discovered, subpoenaed, archived, tracked, and exploited to all sorts of ends. Such a beast has never before existed in the history of culture, but is almost guaranteed to grow exponentially from this day forward. This artifact can tell us extraordinary things about who we are and what we want as a culture. And it has the potential to be abused in equally extraordinary fashion. (Batelle, 2003)*

- *Sabeer Bhatia and Jack Smith – founded Hotmail, the first free internet email service, and conceived an amazing way to market it*

  Fundadores de **Hotmail**, el primer servicio de correo electrónico gratuito, 1996

- *Rick Boyce – credited with pioneering the idea of banner advertising*

  Pionero en la idea de publicidad en ***banners***, 1994

- *Sergey Brin – technology to enhances people's lives and founded Google*

  Cofundador de **Google** 1998

- *Andrei Broder – three types of web queries (informational, navigational, transactional)*

  Identifica tres **tipos de búsqueda** en la web: informacional, navegacional y transaccional. 2002

- *Nico Brooks – conceived of the 'conversion potential' in sponsored search*

  Introduce la idea de ***conversion potential*** en búsquedas patrocinadas.

- *Clay Davis – provided me the idea for using 'the 7 words you can't say on television'*

- *Rufus Evison - The hotel problem*

  Metáfora para explicar el primer problema con el que nos encontramos a la hora de enfrentarnos a un studio de analítica web.

- *Gord Hotchkiss – conceived the idea of the Google Triangle based on eyetracking studies*

  Concibió la idea del *Triángulo de Google* basado en estudios de **eyetracking** 2005

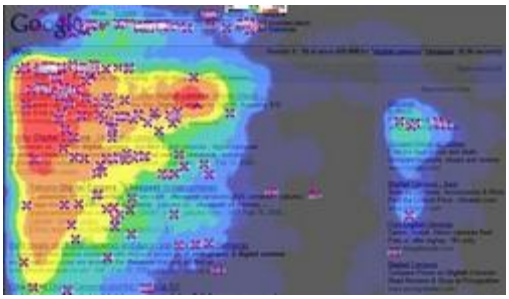


Figura 12. Eyetraking. Fuente (Jansen, 2011)

El área de la página de resultados de Google por la que más pasa la vista según estudios de eyetracking

- *Brad Geddes – explains the three entities in web searching*

  Explica las tres entidades en búsquedas por Internet

- *Salar Kamangar and Eric Veach – developed Google AdWords, which provides ~90+% of Google's revenue*

  Desarrolladores de **AdWords** de Google, que proporciona aproximadamente el 90% de los ingresos de Google. 2000

- *Marissa Mayer – the Google business model*

  El **modelo de negocio** de Google

- *Jamie Murphy, Daehee Park, and Lee Hunter – the Google Online Marketing Challenge*

- *Larry Page – the concept of ads being relevant and non-annoying for web searching*

  Introduce el concepto de anuncios que son pertinentes y no molestos para búsquedas por Internet

- *Jan Pedersen – web searching behaviors*

  Web, buscando **comportamientos**

- *David Penniman – developed standard methods of log analysis*

  Desarrollo de métodos estándar de **análisis de registros**

- *Pete Pirolli – information foraging theory*

  Teoría de la **búsqueda de información**

- *Marc Resnick – user reaction to sponsored links*

  Reacción de los usuarios a los enlaces patrocinados

- *Dan Rose – user intent in web searching*

  La **intención** del usuario en búsquedas por Internet

- *Tefko Saracevic – the concept of relevance*

  El concepto de **pertinencia**. Concepto que ha revisado en diferentes artículos, a lo largo de toda su carrera, desde 1975 hasta 2008

- *Eric Schmidt – Google being in the auction business*
- *Amanda Spink – web searching behavior*
- *Danny Sullivan – commentary on web searching*
- *Mimi Zhang – search engine branding and viral marketing*

Jansen y Pink hablan de las búsquedas en internet buscando patrones de comportamiento de los usuarios: Longitud de la cadena de búsqueda, refinamiento de las búsquedas (Jansen y Pink, 2001),… trabajo que han seguido desarrollando en otros posteriores, en los que se comparan las interacciones que ocurren entre los usuarios y los motores de búsquedas desde la perspectiva de la duración de la sesión, la duración de la consulta, la complejidad de la consulta, y acceso al contenido listado por cada motor de búsqueda Web. (Jansen y Pink, 2004).

También han publicado artículos referidos a los documentos recuperados y su visualización. (Jansen y Pink, 2003). Se profundiza más en la pertinencia de los resultados y cómo afecta al número de páginas de respuestas que consultan los usuarios.

En 2010 Batelle actualiza su *Database of intentions* reconociendo que se equivocó al suponer que la base de datos de las intenciones de la que hablaba en 2003 se crea solo a través de nuestras interacciones con las búsquedas tradicionales (Batelle, 2010). Con la evolución y aparición de nuevos campos de acción de los usuarios de internet, esa base de intenciones necesariamente se ha ampliado, y está formada, según *Batelle* por:

1. Las búsquedas, que ya han sido explicadas con anterioridad.
2. El entramado social que creamos con nuestros contactos en algunas redes sociales, quién soy y a quién conozco,...

3. Una tercera señal que configura nuestra base de datos será el estado, los comentarios personales de lo que nos parezca digno de mención en cada momento.
4. El *check-in* o ¿dónde estoy? que es un aspecto que se acentúa más con el uso de dispositivos móviles.

Resaltamos esta actualización de Batelle para destacar que efectivamente las búsquedas nos van a dar, sí o sí, información real sobre nuestros usuarios. Pero también para hacer hincapié en el hecho de que el usuario desarrolla diferentes acciones en la web, y que cada una de las acciones que realiza nos puede dar información importante y a tener en cuenta durante la toma de decisiones de nuestros proyectos.

3.3.2. Intención de búsqueda

Aunque hay estudios anteriores sobre intención de búsqueda, es Broder (2002) el que propone la clasificación de las búsquedas, actualmente aceptada, en tres tipos: transaccional, navegacional e informativa. Broder utiliza dos métodos para determinar la intención de búsqueda, encuestas a usuarios y análisis manual de un archivo de registro de Alta-Vista.

Lee et al. (2005) proponen un modelo automático de identificación de la intención de búsqueda sin una respuesta explícita del usuario, basándose en el comportamiento previo del usuario antes de hacer la búsqueda y los clicks que realiza después sobre los resultados de las búsquedas.

Baeza_Yates *et al.*(2006) ofrecen un marco para la identificación automática de la intención de búsqueda, basado en el análisis de *logs*. La identificación se hace desde dos perspectivas, los objetivos del usuario, y la categoría en la que se enmarcan esos objetivos.

Dando por hecho que la búsqueda es una representación de una necesidad, implícitamente se representan una serie de factores que serán los que se traten de identificar para poder identificar cuál es la intención de búsqueda.

Jansen *et al.* (2007) proponen un algoritmo de búsqueda basado en la observación de 5 millones de búsquedas, de 7 *logs*, archivos de registro, de 3 buscadores

diferentes. Se han identificado características según la intención de búsqueda, con las que propone la implementación del algoritmo.

Jansen y Booth (2010) investigan, siguiendo la propuesta de Baeza_Yates *et al.*(2006) una metodología para clasificar automáticamente las búsquedas por intención y por tema.

Todos estos estudios sobre intención de búsqueda tienen como objetivo mejorar el rendimiento del buscador, con la identificación más eficaz de las consultas.

3.3.3. Propuesta

En este trabajo proponemos un estudio de las búsquedas partiendo de la premisa de que estudiamos formas de describir contenido cuando el usuario interroga al buscador.

Se aborda este estudio siguiendo la filosofía de trabajo planteada en las publicaciones sobre búsquedas web, recopilando datos de búsquedas y estudiando aspectos concretos de las mismas para contestar unas preguntas previas. En nuestro caso recopilaremos datos de búsquedas, desde la perspectiva de la mejora de los sistemas de información relacionados con las bibliotecas.

La elección de un tema concreto entronca con los estudios sobre intención de búsqueda y clasificación de la misma por temas. En nuestro caso partimos ya de un tema concreto, e intentaremos identificar cuál es la intención de búsqueda del usuario.

Además, estudiaremos las búsquedas en comparación con los encabezamientos de materia de las bibliotecas públicas, estudiando el aspecto léxico de las propias búsquedas.

4. HIPÓTESIS

Para definir un perfil de usuario, y averiguar cómo se enfrenta a la tarea de la descripción de contenido, planteamos una serie de hipótesis que nos obligarán a obtener distintos datos para determinar si las hipótesis planteadas son ciertas o falsas.

Hipótesis

1. El usuario habitual de internet ha modificado su conducta, ya no busca igual que hace diez años.

   En los últimos años, con la experiencia adquirida, y teniendo en cuenta que no hablamos de búsquedas generales sino de búsquedas en el entorno de la biblioteca, queremos demostrar que un usuario típico, sin un perfil ni conocimientos específicos, ya no busca igual que hace diez años.

   Tenemos datos de estudios realizados desde 1998 sobre cómo buscan los usuarios en la web. Haciendo un recorrido por los mismos podemos observar si ha habido algún cambio significativo en las conductas de búsquedas.

2. El lenguaje utilizado por los usuarios que utilizan el castellano para realizar sus búsquedas en Google.com, sobre temática bibliotecaria, no coincide con los encabezamientos de materia de las bibliotecas públicas españolas, listado utilizado por los bibliotecarios de las bibliotecas públicas españolas y en castellano.

   Esta hipótesis podríamos considerarla como una segunda parte de la anterior. Volvemos a hablar de la tarea del profesional de la información, porque en el análisis de contenido, en la fase de tratamiento técnico de la cadena documental, el indizador no solo asigna una materia al documento para su identificación en la colección, sino también para su recuperación, en la última fase de la cadena, la de la salida de información.

   Los encabezamientos de materia han sido redactados para una biblioteca pública española de carácter general y de tipo medio. El público de una biblioteca pública podemos equipararlo al del usuario de Google, y al usuario de Library Thing, ya que es una comunidad que no se dedica a la catalogación de libros de una materia determinada.

Estas hipótesis se plantean pensando en que la información que extraigamos ayudará a mejorar un sistema de información de una biblioteca pública. Se trata de poner al usuario en el lugar del profesional que tiene que organizar el contenido, saber cómo describiría un usuario un libro o cualquier otro recurso de información que pueda obtener en nuestra biblioteca.

Para contestar a cada una de estas hipótesis es necesario realizar un estudio en el que valoremos distintos aspectos de las búsquedas.

La intención de este trabajo será generar una metodología de análisis, para la cual realizamos estos estudios. Una vez analizada la información obtenida de cada estudio se hará una propuesta metodológica general de estudio de hábitos de búsqueda.

¿Qué podemos estudiar para responder a las hipótesis planteadas?:

Con la información que tenemos podemos observar una serie de aspectos sobre las búsquedas: aspectos formales e información acerca del contenido.

De los aspectos formales podemos estudiar:

- longitud de la cadena de búsqueda. Número de palabras con las que se realiza una consulta a un buscador. Estos datos los podemos comparar con los de otros estudios. Compararla con datos de estudios anteriores y averiguar si se busca igual que hace diez años.

Estudiando el contenido podemos averiguar:

- Intención de búsqueda, rastreando algunos términos que nos darán pistas sobre el objetivo del usuario
- si se busca por formatos, por fechas,… los subencabezamientos nos pueden ayudar a identificar necesidades de búsqueda. Qué prefieren buscar los usuarios. Y qué buscan que las bibliotecas públicas no contemplan, o por lo menos sus listas de encabezamientos.

5. MATERIAL Y MÉTODOS

5.1. Material de trabajo

Se han consultado distintas fuentes buscando la bibliografía relacionada con el tema de estudio, y para obtener los datos del experimento.

- **Google Adwords**, herramienta comercial de Google a través de la cual hemos obtenido los datos de búsquedas de usuarios.

  Google Adwords es un producto de Google para la creación y publicación de anuncios en las páginas de resultados de Google y en su red de publicitaria. Hemos utilizado Google Adwords porque, a través de su herramienta de búsquedas de palabras clave nos ofrece información en abierto de las consultas que realizan sus usuarios a través de Google. Estamos utilizando datos públicos, que están disponibles para todos los usuarios. Además Google Adwords nos da la información de búsquedas de Google, que es el principal buscador utilizado en el mundo en 2009, con una cuota de mercado superior al 85%. En países como Alemania, España o Reino Unido alcanza una cuota superior a los 90 puntos porcentuales. Aunque Yahoo! se posiciona en segundo lugar y en el contexto mundial alcanza una cuota de mercado superior a los seis puntos, en los países europeos tiene una presencia meramente testimonial, al igual que ocurre con el resto de buscadores. (eespaña 2010, Informe anual sobre el desarrollo de sociedad de la información en España).

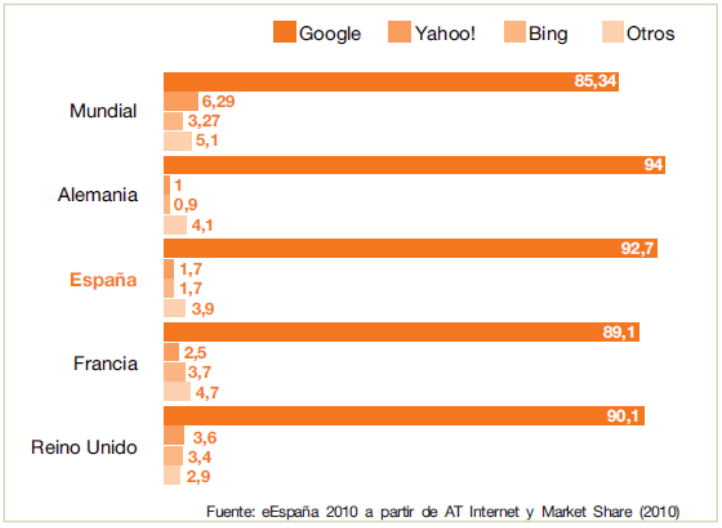


Figura 13. Cuota de mercado de los principales buscadores en 2009, en %

- Catálogos, Bases de datos y portales de información bibliográfica: ACM, Dialnet, ISOC (CSIC), Catálogo UPV, Metabuscador UPV, etc.
- Bibliografía básica y específica sobre el tema
- Lista de Encabezamientos de materia de las Bibliotecas Públicas (LEMBP)
- Microsoft Office VISIO para la realización de esquemas y mapas conceptuales
- Microsoft Office Excel 2007 para analizar los datos y creación de gráficas.
- Microsoft Office Word 2007 para redacción y edición del proyecto.

5.2. Metodología

Se establece un marco comparativo entre los datos proporcionados por *Adwords* y los datos que recoge la Lista de Encabezamientos de Materia de las Bibliotecas Públicas, LEMBP, que es el medio más sencillo y menos formalizado de la biblioteca (Dirección General del Libro y Bibliotecas, 1994) para mediar entre la búsqueda de información sobre un campo determinado de la realidad que realiza un usuario y la información que busca sobre esa materia, o sobre un libro del que solo conoce la materia de que trata.

La LEMBP se genera siguiendo unas pautas: es una lista de términos organizada en dos niveles, encabezamientos y subencabezamientos. Los encabezamientos hacen referencia a las materias, y los subencabezamientos especificarán la materia del encabezamiento.

Se utilizarán en casos diferentes el singular y el plural, se recomienda utilizar un término en vez de otro para asignar una materia a un contenido, y de este modo evitar la sinónima,…

El uso de la LEMBP está basado en unos principios básicos, y nos apoyaremos en ellos para estudiar cómo realiza el usuario las búsquedas.

Esta ha sido la metodología empleada:

1. Búsqueda de artículos relacionados con el tema. Búsqueda bibliográfica (Catálogo UPV, ACM, blogs, Google Scholar), explicado en el apartado *3.1. Recopilación de información*.
2. Primero se realizó una toma de contacto realizando búsquedas en Google Scholar que nos dieron pistas de cuáles eran los autores más relevantes.

3. Exploración de la herramienta Google Adwords. Cómo se obtienen las palabras clave relacionadas con un término y el uso de los filtros de selección.
4. Lectura en profundidad de artículos referidos a aspectos básicos de la búsqueda de información.
5. Recopilación de datos con Google Adwords.
6. Tratamiento de los datos y generación de la lista definitiva de términos con los que vamos a trabajar.
7. Comparación de términos:

   Palabras clave utilizadas para búsquedas con la LEMBP

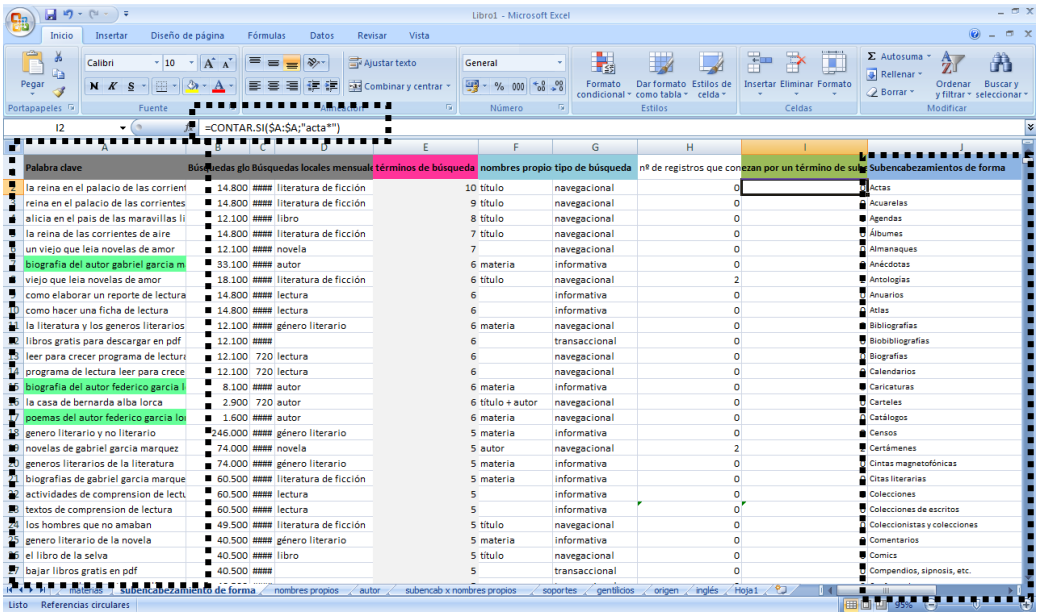


Figura 14. Hoja de cálculo utilizada en este estudio. Fórmula para detectar coincidencias entre encabezamientos y búsquedas

Comparar palabras clave con encabezamientos supone buscar coincidencias de términos.

Para buscar coincidencias entre palabras clave y encabezamientos en una columna tenemos las palabras clave, y en otra los subencabezamientos, en la imagen los de forma.

Con la fórmula CONTAR.SI utilizamos los subencabezamientos como criterio de búsqueda, y las palabras clave como rango dentro del cual queremos contar las celdas. =CONTAR.SI($A:$A;"acta*")

De este modo contabilizamos el número de palabras clave que contienen, por ejemplo, la palabra *libro* en su cadena de búsqueda, o la palabra *acta*, o *atlas*,…

8. Análisis de datos.
9. Conclusiones.

5.3. Datos. Limitaciones. Delimitación del universo y la muestra.

Para llevar a cabo el estudio hemos seleccionado datos de diferentes fuentes:

a. palabras clave de Google Adwords,

b. Encabezamientos y subencabezamientos de materia de la Lista de Encabezamientos de Materia de las Bibliotecas Públicas

5.3.1. Datos

Se han recogido 2 tipos de datos:

Palabras clave utilizadas en las búsquedas	Nº de búsquedas realizadas con las P.C.
772	171.070.500

- **772 palabras clave**,

  ¿Cómo hemos obtenido las palabras clave?

  Hemos seleccionado un total de 12 palabras que serán la base para generar el listado de palabras clave con la que intentaremos averiguar cómo buscan los usuarios en la red, cuando quieren obtener información a través de un buscador.

  Partiendo de la palabra *libro* hemos elegido el resto de términos que generarán palabras clave, siguiendo criterios de relación:

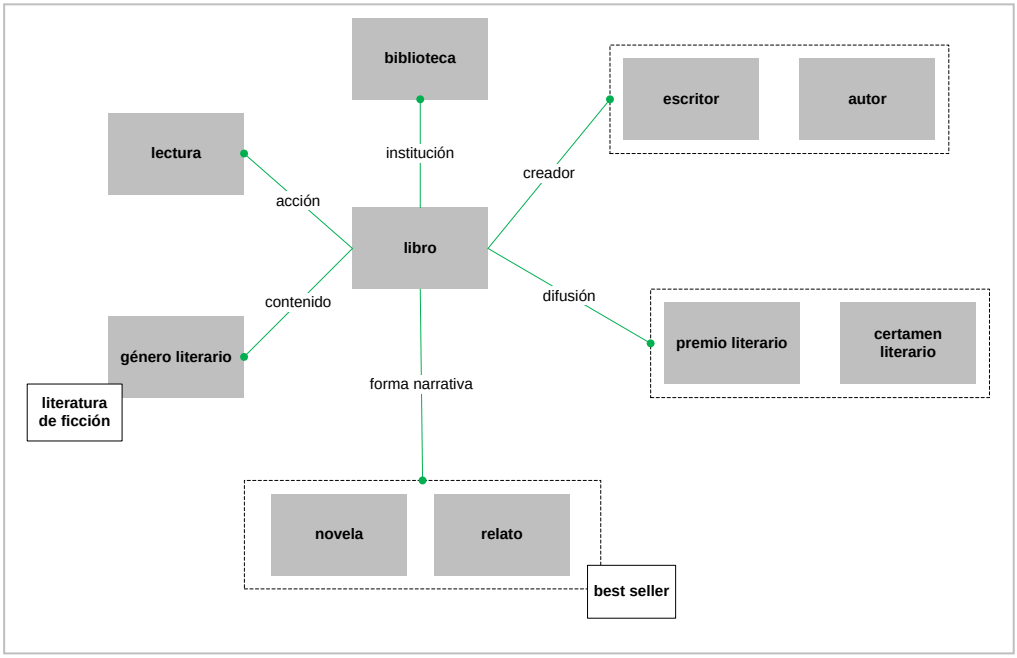


Figura 15. Representación gráfica de la toma de decisiones en elección de palabras clave

Pasos:

1. Elegimos los términos que describen el tema sobre el que queremos trabajar, en este caso el tema son libros y bibliotecas:

   Biblioteca, Lectura, Libro, Género literario, Literatura de ficción, Novela, Relato, *Best Seller*, Escritor, Autor, Premio literario, Certamen literario.

2. Uno a uno buscamos, con la herramienta de Generación de Palabras clave de Google Adwords, las Palabras clave relacionadas con cada uno de los 12 términos de búsqueda.

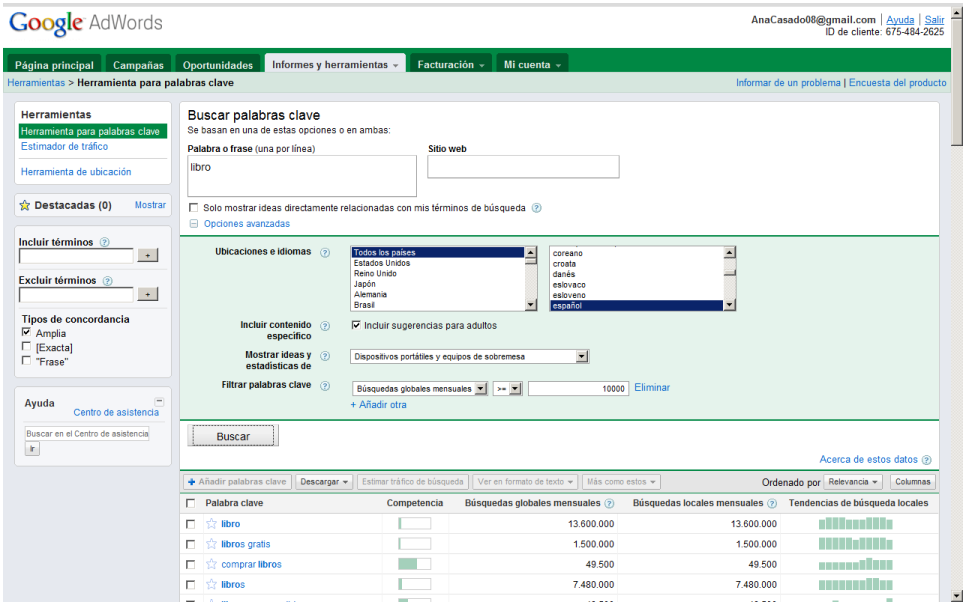


Figura 16. Interfaz Herramienta para buscar palabras clave de Google adwords

a. Escribimos el término de referencia en el cuadro donde pone Palabra o frase

b. En Ubicaciones e idiomas, seleccionamos el idioma “español”

c. Tipo de concordancia: Amplia

d. Desactivamos la opción de “Solo mostrar ideas directamente relacionadas con mis términos de búsqueda”, que restringe los resultados a búsquedas en las que obligatoriamente aparece el término de referencia.

e. En “Incluir contenido específico”, seleccionamos “Incluir contenido para adultos”

f. Especificamos un filtro de palabras clave:

   Búsquedas globales mensuales >= 10.000

g. Pinchamos en “Buscar”

h. Obtenemos el listado con las palabras clave relacionadas. Exportamos los resultados en formato *.csv, Archivo de valores separados por comas de Microsoft Office Excel.

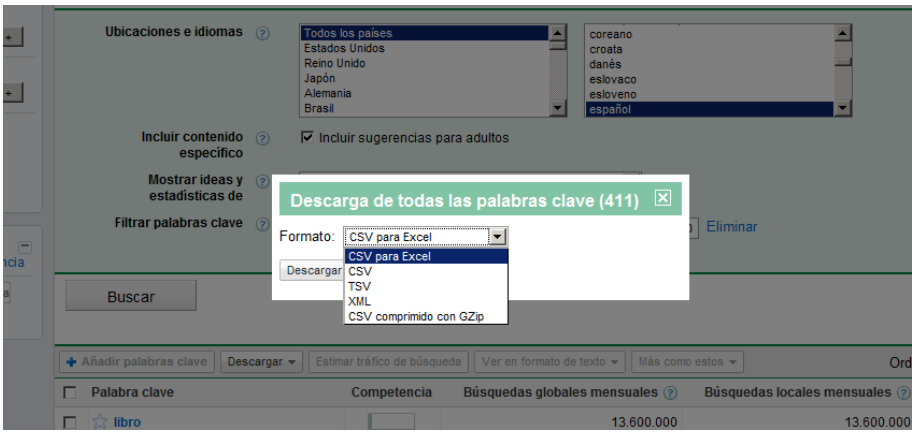


Figura 17. Cuadro de diálogo para elegir formato de archivo de descarga de datos de Palabras clave. Google Adwords

i. Repetimos la operación con los siguientes términos de búsqueda.

Para elegir las palabras clave definitivas:

3. Tenemos un listado por cada término de referencia. De cada listado eliminamos:
   a. Las columnas de datos que no hagan referencia a “Palabra Clave”, “Búsquedas globales mensuales” y “Búsquedas locales mensuales”
   b. Los duplicados con menos búsquedas asociadas.
   c. Del listado relacionado con el término de referencia “novela” eliminamos los términos que hacen referencia a telenovelas.
   d. Guardamos cada archivo como archivo *.xlsx, Hoja de cálculo de Microsoft Office Excel.
4. Unificamos todos los listados.
5. Eliminamos las palabras clave duplicadas.

De este modo obtenemos el listado definitivo de palabras clave de búsqueda, que contiene **772** palabras clave, y entre todas suman una media de 171.070.500 de búsquedas globales mensuales, y 33.272.010 de búsquedas locales mensuales

- **171.070.500 búsquedas**

  realizadas con las 772 palabras clave seleccionadas.

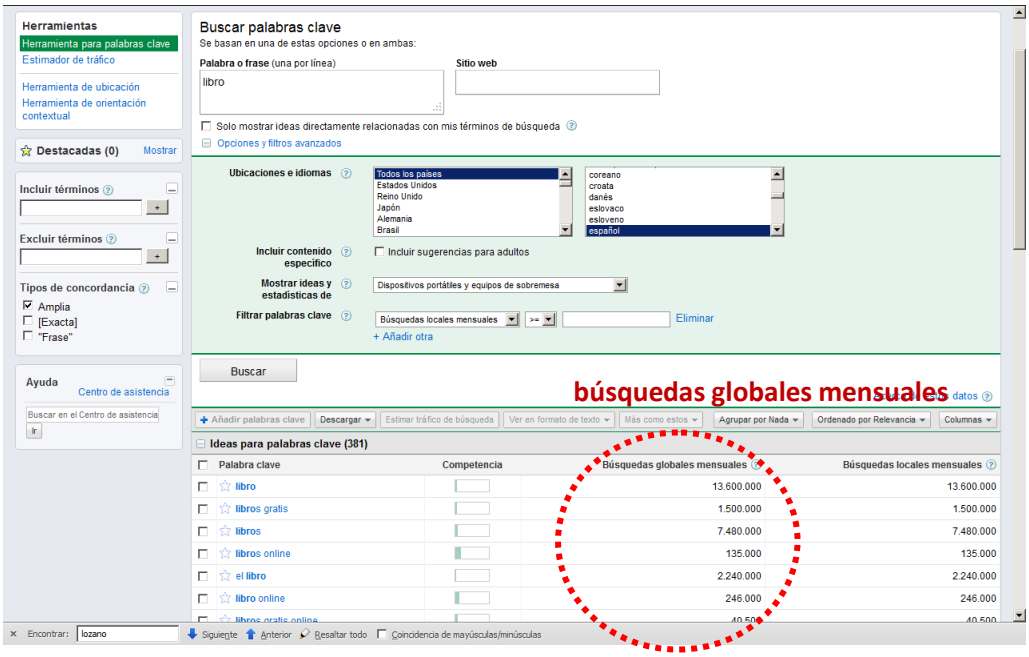


Figura 18. Interfaz de búsqueda de palabras de Google Adwords.
Búsquedas globales mensuales destacadas

Este dato lo hemos obtenido sumando el dato que nos proporciona Google Adwords bajo el epígrafe "Búsquedas globales mensuales", y es el *promedio anual aproximado de consultas de los usuarios que incluyen la palabra clave en la búsqueda en Google* (Google Adwords)

5.3.2. Limitaciones y Decisiones

Limitaciones en la obtención de los datos del estudio, y decisiones en el tratamiento de los datos:

5.3.2.1. LIMITACIONES

- Adwords solo nos ofrece información de los últimos 12 meses. Los datos definitivos de las búsquedas se obtuvieron en diciembre de 2010 y corresponden a las búsquedas realizadas entre diciembre de 2009 y noviembre de 2010, ambos incluidos.
- No trabajamos con archivos log (archivos de registro de transacciones de buscadores), en los que quedan registrados todos los datos de las sesiones, por lo tanto no podemos manejar datos relacionados con perfiles de usuario. ni sobre refinamiento de búsquedas. Con los datos que tenemos no podemos

saber si las palabras clave que nos da la aplicación son búsquedas únicas o modificaciones de una búsqueda anterior. Es decir, no tenemos información sobre sesiones.

- Estamos utilizando los datos que nos proporciona gratuitamente Google Adwords. Pero nosotros no tenemos acceso a los datos en bruto. Antes de poner a disposición del usuario los datos de búsqueda y etiquetado, ha habido un redondeo.

5.3.2.2. DECISIONES

- De las palabras clave vamos a analizar los siguientes aspectos, que son medibles y cuantificables:

- Cómo se construye la cadena de búsqueda. Longitud de la cadena de búsqueda
- Las palabras clave, ¿se aproximan al lenguaje controlado de la LEMBP?

  Se comparan las palabras clave con los encabezamientos de la LEMBP, para establecer cuál es la distancia entre el lenguaje de descripción de contenidos que se utiliza en una biblioteca y el que utiliza el usuario cuando quiere acceder al documento.

  a. Consideraremos como materia, palabras clave como:

     Libros de anatomía por *anatomía*

     Libros de biología por *biología*

     Libros medicina por *medicina*

     Entendiendo “libros de” y “libros” como partícula de contextualización de la búsqueda.

  b. Se aceptarán expresiones de la forma:

     modernismo autores y *autores modernistas*

     libros de Dan Brown y *gabriel garcia marquez novelas*

  c. Se han contabilizado tanto las formas recomendadas como las no recomendadas:

     Certamen poesía 2010 y *concurso poesía 2010*

d. Indistintamente se han aceptado las formas en singular y plural, aunque en la lista puedan tener connotaciones distintas.

e. La LEMBP utiliza algunos subencabezamientos de forma para describir el soporte del material que describe. Es el caso de términos como: Videocassetes, Videodiscos, Fotografías, Grabaciones en vídeo, Grabaciones sonoras o Grabados.

   Teniendo en cuenta que una de las ideas básicas del LEMBP es *Partir de la idea de un esquema ideal de colección para biblioteca pública de nuestro tiempo que quiera ser reflejo fiel de la cultura actual, y en las normas de la Lista*, en el apartado de Finalidad leemos *Sería ingenuo pensar que esta Lista ofrece todas las soluciones, pero sería suicida creer que, desde ahora, puede prescindirse de ella. Sobre su pauta pueden construirse otras generales con mayor número de encabezamientos y un número indefinido para materias especiales que no llegarían a ser todavía un "thesaurus"* (Dirección General del Libro y Bibliotecas, 1994*)*, lo que nos indica que no es una lista cerrada y nos invita a ampliarla según nuestras necesidades, se han incluido en el recuento de los Subencabezamientos de forma los siguientes términos, que no se contemplan en la Lista original: ebook, ereader, pdf, dvd, audiolibro, libros digitales, libros electrónicos, on line.

   Esto es, en el recuento de subencabezamientos de forma se han incluido las palabras clave que incluían los términos ebook, ereader, pdf, dvd, audiolibro, libros digitales, libros electrónicos y on line.

f. Siguiendo el espíritu de la LEMBP podríamos considerar *libros digitales*, *audiolibros*, y *ereaders* como materias si estuviésemos buscado documentos cuyo contenido tratase sobre alguno de estos soportes. Se ha tomado la decisión de considerarlos subencabezamientos, en primer lugar porque son términos en plural, y para ser materias deberían estar expresados en singular y en segundo lugar porque es más factible que el usuario realice esta búsqueda considerando estas palabras clave como soportes de contenido de otro tipo.

g. La LEMBP contempla 6 tipos de subencabezamiento:
   a. de forma

b. bajo nombres de países
c. bajo nombres de ciudades
d. para la historia de España
e. para la historia de los países iberoamericanos
f. para nombres de persona

Para este estudio no vamos a utilizar los relacionados con la historia de España y de los países iberoamericanos, ya que consideramos que se refieren a un tema muy específico.

Pero sí que se tendrán en cuenta los subencabezamientos bajo nombres de países y ciudades, que nos pueden ayudar a determinar búsquedas geográficas, que no hacen referencia necesariamente a un tema específico, y también los subencabezamientos de forma y bajo nombres de persona.

- Para decidir la Intención de cada búsqueda seguiremos los siguientes criterios, que nos permitirán clasificar las búsquedas como:
  - Navegacional: En este grupo se han incluido las búsquedas del tipo bibliotecas públicas Madrid, editorial Santillana, novelas García Márquez,…
    - Aquellas búsquedas que explícitamente hacen referencia a una institución, negocio o nombre propio
  - Transaccional: El objetivo del usuario, no es encontrar o localizar contenidos informativos, sino realizar transacciones y acciones de intercambio.
    - Búsquedas que contienen los términos gratis, descargar, comprar, online.
  - Informativa: El usuario busca ampliar su conocimiento sobre algún tema, aprender, leer, obtener un dato concreto, informarse, etc.
    - Búsquedas que no siguen los criterios de clasificación de búsquedas transaccionales o navegacionales.

5.4. Diseño de los experimentos

5.4.1. Obtención de datos para el estudio.

Se generará un grupo de palabras clave formado por términos de búsqueda empleados en Google.es. Los términos se recopilaran a través de Google Adwords.

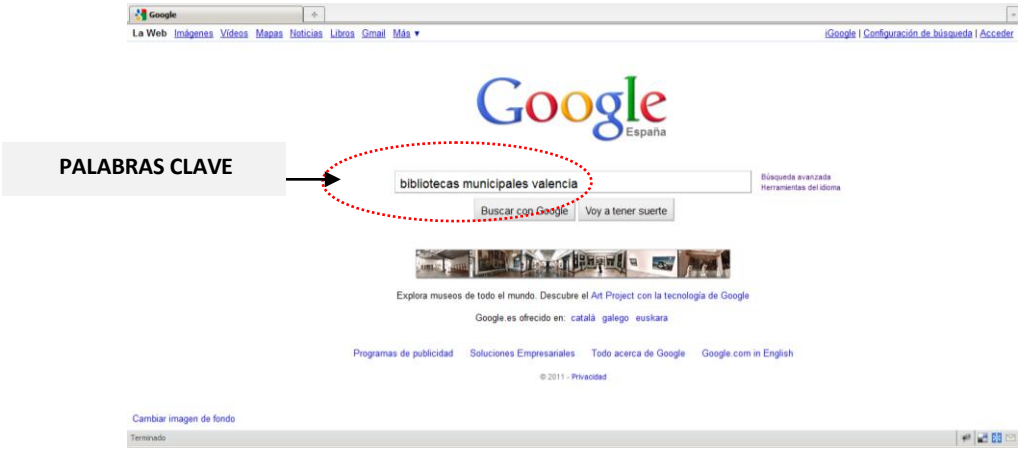


Figura 19. Interfaz Página principal Google.es

El análisis de estas palabras clave, nos ofrece información de cómo buscan los usuarios a través de Google, cómo se buscan títulos famosos, autores famosos, cómo la gente busca autores por "*escritores chilenos*", en lugar de "escritores-Chile" al estilo de los encabezamientos de materia…

5.4.2. Extracción de datos

Extracción de los datos más significativos, que nos ayudarán a responder a las hipótesis planteadas:

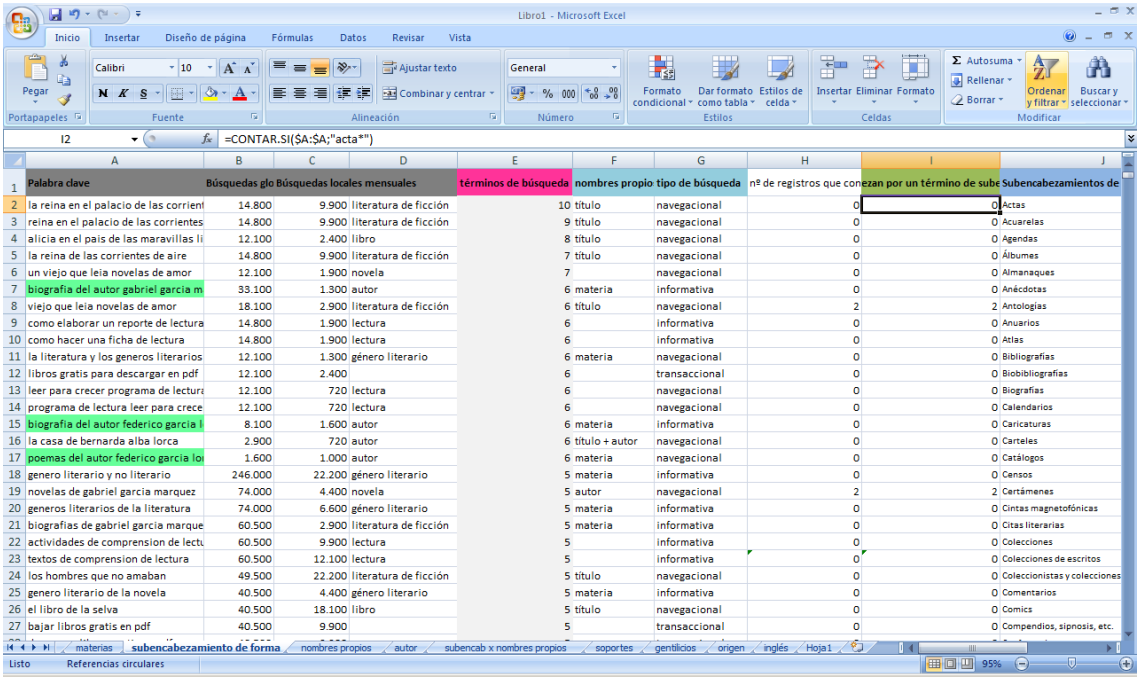
El recuento y la extracción de datos han sido básicamente manuales. Se han automatizado las operaciones más costosas utilizando fórmulas de Excel y tablas dinámicas, con las que se han extraído datos combinados. Se han tomado decisiones organizativas y se han contabilizado datos como:

a. longitud de la cadena de búsqueda. Número de palabras que componen cada palabra clave.

Fórmula de Excel:

=SI(LARGO(ESPACIOS(A2))=0;0;LARGO(ESPACIOS(A2))-LARGO(SUSTITUIR(A2;" ";""))+1)

b. número de palabras clave que contienen encabezamientos o subencabezamientos de materia.
   Fórmula de Excel: =CONTAR.SI($A:$A;"acta*")
c. asignación de un tipo de intención de búsqueda a cada palabra clave.
   - Navegacional: Aquellas palabras clave que explícitamente hacen referencia a una institución, negocio o nombre propio.
   - Transaccional: palabras clave que contienen los términos gratis, descargar, comprar, online. Utilizando la opción de búsqueda de Excel (ctr+B)
     No hemos incluido en nuestro recuento de transaccionales búsquedas que contengan términos como imágenes, vídeo, pdf, como Jansen *et al*. en su estudio de determinación de intención de búsqueda (2007), porque son términos que describen formatos, pero no necesariamente implican una transacción.
   - Informativa: palabras clave que no siguen los criterios de clasificación de las transaccionales o navegacionales.
d. recuento manual de palabras clave que contienen nombres propios de persona, de instituciones, títulos de obras o gentilicios.
e. recuento manual de palabras clave en inglés.



	Palabra clave	Búsquedas glo	Búsquedas locales mensuales		términos de búsqueda	nombres propio	tipo de búsqueda	nº de registros que comienzan por un término de sube		Subencabezamientos de
2	la reina en el palacio de las corrient	14.800	9.900	literatura de ficción	10	título	navegacional	0	0	Actas
3	reina en el palacio de las corrientes	14.800	9.900	literatura de ficción	9	título	navegacional	0	0	Acuarelas
4	alicia en el pais de las maravillas li	12.100	2.400	libro	8	título	navegacional	0	0	Agendas
5	la reina de las corrientes de aire	14.800	9.900	literatura de ficción	7	título	navegacional	0	0	Álbumes
6	un viejo que leia novelas de amor	12.100	1.900	novela	7		navegacional	0	0	Almanaques
7	biografia del autor gabriel garcia m	33.100	1.300	autor	6	materia	informativa	0	0	Anécdotas
8	viejo que leia novelas de amor	18.100	2.900	literatura de ficción	6	título	navegacional	2	2	Antologías
9	como elaborar un reporte de lectura	14.800	1.900	lectura	6		informativa	0	0	Anuarios
10	como hacer una ficha de lectura	14.800	1.900	lectura	6		informativa	0	0	Atlas
11	la literatura y los generos literarios	12.100	1.300	género literario	6	materia	navegacional	0	0	Bibliografías
12	libros gratis para descargar en pdf	12.100	2.400		6		transaccional	0	0	Biobibliografías
13	leer para crecer programa de lectura	12.100	720	lectura	6		navegacional	0	0	Biografías
14	programa de lectura leer para crece	12.100	720	lectura	6		navegacional	0	0	Calendarios
15	biografia del autor federico garcia l	8.100	1.600	autor	6	materia	informativa	0	0	Caricaturas
16	la casa de bernarda alba lorca	2.900	720	autor	6	título + autor	navegacional	0	0	Carteles
17	poemas del autor federico garcia lo	1.600	1.000	autor	6	materia	navegacional	0	0	Catálogos
18	genero literario y no literario	246.000	22.200	género literario	5	materia	informativa	0	0	Censos
19	novelas de gabriel garcia marquez	74.000	4.400	novela	5	autor	navegacional	2	2	Certámenes
20	generos literarios de la literatura	74.000	6.600	género literario	5	materia	informativa	0	0	Cintas magnetofónicas
21	biografias de gabriel garcia marque	60.500	2.900	literatura de ficción	5	materia	informativa	0	0	Citas literarias
22	actividades de comprension de lectu	60.500	9.900	lectura	5		informativa	0	0	Colecciones
23	textos de comprension de lectura	60.500	12.100	lectura	5		informativa	0	0	Colecciones de escritos
24	los hombres que no amaban	49.500	22.200	literatura de ficción	5	título	navegacional	0	0	Coleccionistas y colecciones
25	genero literario de la novela	40.500	4.400	género literario	5	materia	informativa	0	0	Comentarios
26	el libro de la selva	40.500	18.100	libro	5	título	navegacional	0	0	Comics
27	bajar libros gratis en pdf	40.500	9.900		5		transaccional	0	0	Compendios, sipnosis, etc.

Figura 20. Hoja de cálculo. Fórmula recuento de palabras

f. generación de gráficas combinando los datos extraídos.

6. RESULTADOS Y ANÁLISIS

6.1. Datos de Palabras Clave y Búsquedas

A continuación se expone la información obtenida a partir de las palabras clave y las búsquedas recopiladas con Google Adwords.

Recordamos que cuando hablamos de "palabras clave" nos referimos a las palabras, expresiones, término o frases que hemos escrito en el buscador para realizar una búsqueda.

A pesar de la confusión que pueda generar, se ha utilizado la expresión "Palabra Clave" porque es el término que utiliza Google Adwords para referirse a la expresión de búsqueda, y tras valorar las distintas opciones que teníamos, y observando que con todas había cierta ambigüedad decidimos mantener la expresión de la herramienta que hemos utilizado para la extracción de datos.

6.1.1. Longitud de la Palabra clave.

Trabajamos con 772 palabras clave con las que se han generado un total de 171.070.500 búsquedas.

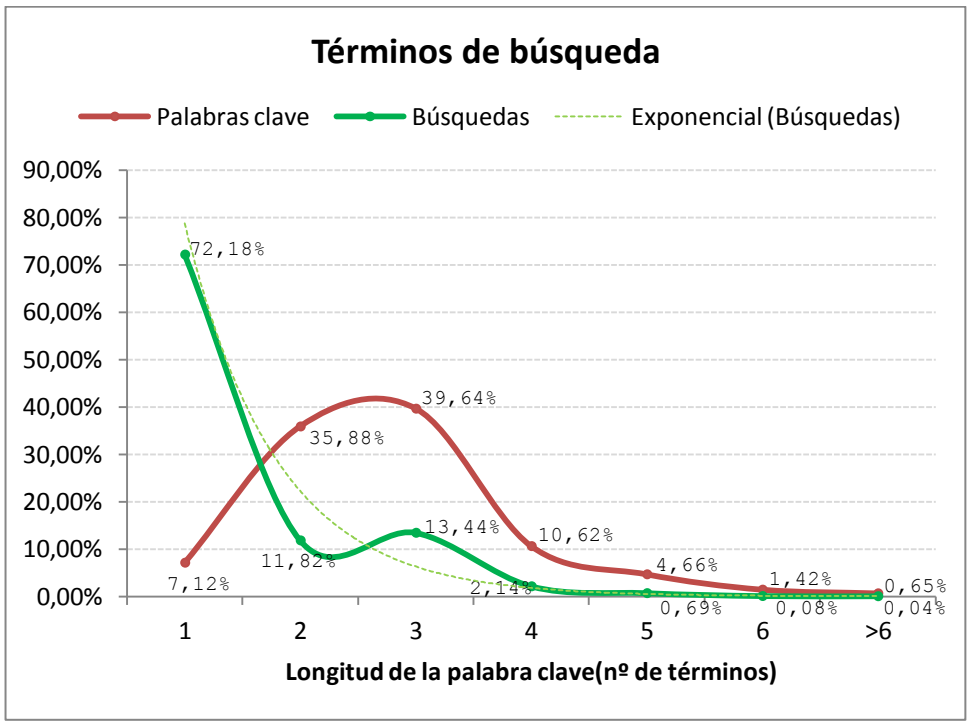


Figura 21. Relación entre longitud de palabra clave y volumen de búsquedas generadas

Con la longitud vamos a averiguar cuántas palabras se utilizan para realizar una búsqueda.

Las palabras clave de longitud 1 término son las más utilizadas. No llegan al 10% del total de las palabras clave y generan el 72,18% de las búsquedas. Dicho de otra manera, el 72,18% de las veces que buscamos algo a través del buscador Google utilizamos una sola palabra, y no manejamos una gran variedad de términos.

Echando un vistazo a los resultados de los estudios de referencia, que no son sobre búsquedas relacionadas con temas específicos, obtenemos algunas diferencias. La primera es el alto porcentaje de búsquedas que hemos obtenido que utilizan solo 1 término. En los estudios de referencia tiene un alto porcentaje de búsquedas, pero siempre rondan un 30%, y en algunos casos son las de 2 términos las más comunes. Nuestros resultados arrojan unos datos significativamente distintos. El 72,18% de las búsquedas se realizan utilizando un solo término. Esto se debe a que entre nuestras palabras clave tenemos algunos términos muy genéricos:

historia, *libro*, *concursos*, *ebook*, *novela*, *cuentos*, *biblioteca*, *poesia*, *literatura*, *leer*.

Estas 10 palabras clave, de longitud 1 término, suman el 49,37% de las búsquedas totales de este estudio. *historia*, el término más utilizado, hace referencia a una materia, y hay otros que buscan información relacionada con soportes como *ebook*, *libro*, o tipologías, *cuentos*, *novela*, *poesía*,…

……

Resultados de otros estudios:

Jansen, B. J., Spink, A., Bateman, J., & Saracevic, T., (1998):

Terms in query	Number of queries	% of all queries
10	185	0.36%
9	125	0.24%
8	224	0.44%
7	484	0.94%
6	617	1%
5	2.158	4%
4	3.789	7%
3	9.242	18%
2	**16.191**	31%
1	**15.854**	31%
0	2.584	5%

B.J. Jansen, A. Spink (2006) :

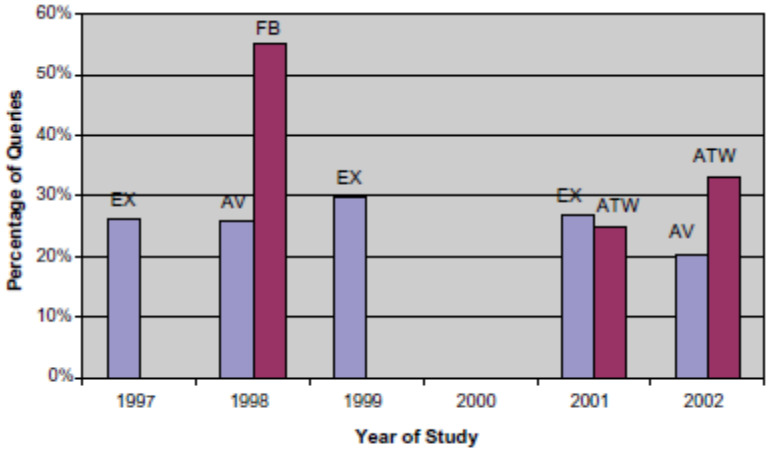


Fig. 2. Percentage of one-term queries.

For the US-based Web-search engines the percentage of one-term queries is holding steady, within a range of 20–29% of all queries. Using data from 1999 onward, the trend with US-based Web-search engines appears to be of one-term queries declining as a percentage of all queries, dropping from 30% to 20%.

For the Europe-based Web-search-engine users, the trend appears to be one of little change, although there is a spike in 2002 with AlltheWeb.com users. Otherwise, we see a percentage of one-term queries on these European-based Web-search engines within a range of about 25–35%, excluding the 1998 Fireball study.

Spink, A., Wolfram, D., Jansen, J., Saracevic, T. (2001):

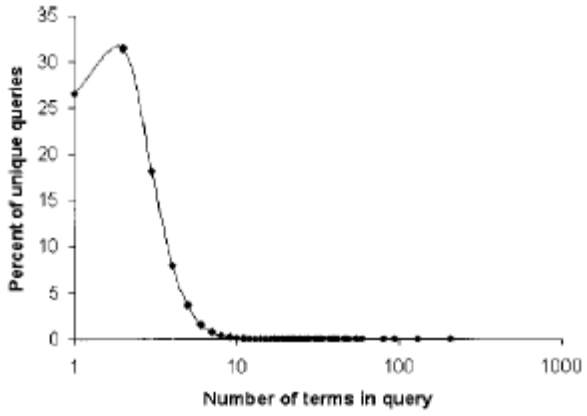


FIG. 4. Number of terms appearing in each unique query. The figure does not include queries containing zero terms, which represent 9.7% of all queries.

The logarithmic scaling does not allow us to include the 9.7% of all queries submitted with zero terms. Web queries are generally short. Some 26.6% of queries had one term only, 31.5% had two terms, and 18.2% had 3 terms. Thus, close to 60% of all queries

had one or two terms, with most of them having the "magical" search length of two terms.

Less than 1.8% of the queries had more than seven terms.

.....

El 11,70% de las búsquedas se formulan con 2 términos y 13,25% utilizando 3 términos, ej. *cuentos para niños*.

La razón por la que en nuestro estudio es más común encontrar búsquedas con 3 términos en vez de 2, es el propio idioma de búsqueda. El 52,77% de las búsquedas que se realizan utilizando 3 términos tienen la estructura:

término + de + término ej. **poemas de amor**

que en inglés equivaldría a una estructura del tipo:

término + término ej. **love poems**

Esta equivalencia coincide en todas nuestras palabras clave con la estructura "término + de + término".

La misma búsqueda que en inglés utiliza 2 términos requiere 3 en castellano. Por lo tanto no podemos contrastar estos resultados con los de otros estudios que se han realizado en inglés.

Las búsquedas con palabras clave largas, a partir de 4 términos, son menos frecuentes, y representan menos del 3% del total de búsquedas. Aun así, no podemos descartarlas ni dejarlas pasar sin tenerlas en cuenta, porque aunque el porcentaje es muy bajo estamos hablando de 4.897.700 búsquedas.

Las palabras clave estudiadas, en general, son cortas, y a partir de 3 términos el volumen de búsquedas se reduce drásticamente.

En general, un pequeño número de términos se utilizan con frecuencia muy alta, (las palabras clave compuestas por un solo término), mientras que hay un gran número términos con una frecuencia de uso muy baja.

No hay grandes diferencias en este aspecto respecto a los resultados de búsqueda de estudios anteriores.

6.1.2. Uso de operadores booleanos

No tenemos datos sobre las sesiones de búsqueda, pero hay un dato que puede avalar nuestros resultados, ninguna búsqueda ha utilizado operadores booleanos:

Operador booleano	Palabras clave	Búsquedas
AND	0	0
NOT	0	0
OR	0	0
Y	20	822.600
O	0	0
NO	3	541.500

Figura 22. Uso de operadores booleanos

¿Por qué decimos que ninguna búsqueda ha utilizado operadores booleanos si tenemos 20 palabras clave que han utilizado "Y" y 2 que han utilizado "NO"?

Porque en ningún caso se han utilizado booleanos en inglés y los que se han utilizado en español aparecen en expresiones en lenguaje natural, sin estructura de búsqueda booleana.

Son búsquedas del tipo:

los hombres que no amaban o *genero literario y no literario*

En las que la partícula ***no*** no es excluyente, aparece en el contexto de la frase, en lenguaje natural y no se utiliza como operador booleano.

6.1.3. Intención de búsqueda

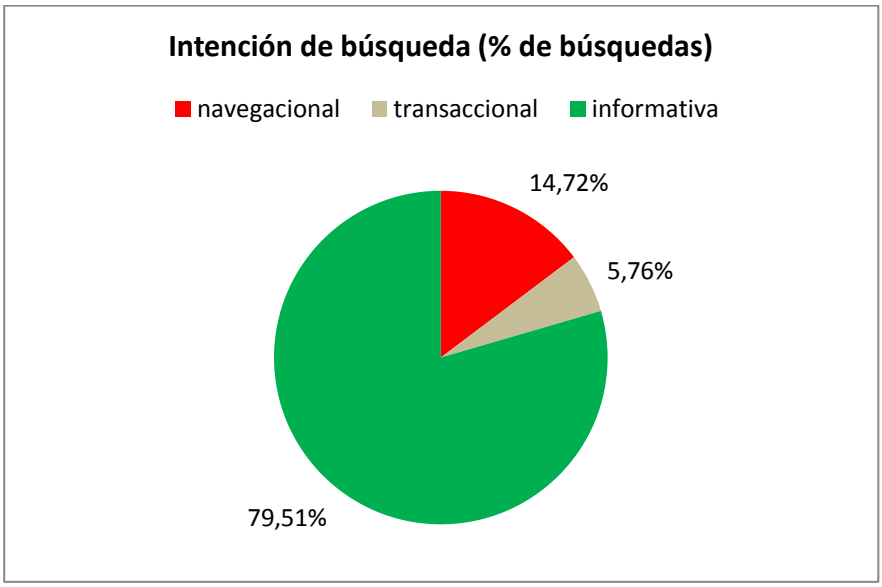


Figura 23. Intención de búsqueda. Datos sobre el volumen de búsquedas

La intención de búsqueda nos habla de las necesidades del usuario, de qué busca en el sentido de la acción, si quiere informase, realizar una transacción, y del nivel de conocimiento del tema sobre el que se quiere informar. Si busca datos específicos, información general,...

De los datos que hemos obtenido se desprende que:

Ese 66,28 % de palabras clave se utilizan en el 79,51 % de las **búsquedas**, que son informativas. Un 5,76 % de las búsquedas son transaccionales, y el resto, el 14,72 % son búsquedas navegacionales.

Esto quiere decir que casi el 80% de nuestras búsquedas se realizan buscando información general sobre un tema. Entre las búsquedas informativas encontramos algunas como: *la literatura contemporánea*, *autores de novelas*, *autores del realismo*.

Solo un 5,76 % de las **búsquedas** relacionadas con los libros son transaccionales, y son del tipo *descargas de libros*, *compras por internet*, *libros on line*, *libros gratis pdf* .

Las búsquedas navegacionales son las que buscan una página o una información concreta. Si tenemos un porcentaje muy alto de búsqueda navegacionales interpretamos que nuestros usuarios ya tienen información previa, y lo que

quieren es ampliar la información sobre algo que ya les es conocido. Podríamos hablar en este caso de usuarios especializados.

Si incorporamos nuestros datos a la tabla de resultados de nivel 1 de Jansen y Booth, (2010, 4288) observamos que:

	Informativa	**Navegacional**	**Transaccional**
Auto	**81.2%**	15.8%	3.0%
Business	47.4%	**51.9%**	0.7%
Computing	**60.5%**	11.8%	**27.7%**
Entertainment	**79.7%**	6.1%	14.2%
Games	65.5%	9.7%	**24.8%**
Health	**89.6%**	8.9%	1.4%
Holiday	48.3%	**50.8%**	0.9%
Home	**60.9%**	21.0%	18.1%
Libros	**79,51%**	14,72%	5,76%
News	**50.9%**	35.1**%**	14.0%
Organization	25.0%	**72.1%**	2.9%
Other	**55.6%**	26.1%	18.3%
Place	**62.9%**	31.1%	6.0%
Porn	11.6%	26.1%	**62.3%**
Research	**51.3%**	32.9%	15.8%
Shopping	33.4%	31.7%	**35.0%**
Sports	**51.7%**	30.2%	18.1%
Travel	47.4%	41.9%	10.7%
URL	0.1%	99.2%	0.7%
Average	51.3%	33.5%	15.3%
Sd Dev	22.9%	23.7%	15.5%
Max	89.6%	99.2%	62.3%
Min	0.1%	6.1%	0.7%

Figura 24. Datos actualizados del estudio de Jansen y Booth con la incorporación de datos de nuestro estudio

Nuestro tema está relacionado con búsquedas mayoritariamente informativas, y obtenemos datos similares a los que se obtienen a las temáticas de coches, salud y entretenimiento. Se busca ampliar información, pero sin buscar un dato específico, y las transacciones son menos demandadas que en otros campos, como juegos y porno.

Los usuarios que se interesan por el tema de los libros buscan en la web más información relacionada con contenido que con descargas o compra de libros.

6.1.4. Longitud de la palabra clave e Intención de búsqueda

La figura 26 muestra el volumen de búsquedas navegacionales, transaccionales e informativas, en relación con la longitud de la palabra clave:

Longitud palabra clave (nº de términos)	navegacional	transaccional	informativa	
1	**7,76%**	0	**64,42%**	72,18%
2	3,29%	2,46%	5,96%	11,70%
3	2,51%	**2,61%**	8,14%	13,25%
4	0,89%	0,64%	0,52%	2,05%
5	0,20%	0,05%	0,44%	0,69%
6	0,03%	0,01%	0,04%	0,08%
>6	0,04%	0	0	0,04%
	14,72%	5,76%	79,51%	TOTALES %

Figura 25. Longitud de palabra clave vs. Intención de búsqueda. Tabla relacional

A primera vista podemos destacar que no hemos registrado búsquedas transaccionales e informativas de más de 6 términos, y tampoco transaccionales de 1 término.

Las búsquedas informativas de 1 término son las más frecuentes, y también que hay una disminución sustancial del volumen de búsquedas utilizando más de 3 términos.

Cuando utilizamos 1, 2 o 3 términos para realizar una búsqueda, mayoritariamente es informativa.

Para las búsquedas informativas se utiliza una mayor variedad de "fórmulas", mientras que para las transaccionales solo utilizamos el 7,49 % de las palabras clave, la mayoría de una longitud de 3 palabras. Esto se traduce en una menor variedad de enunciados a la hora de realizar búsquedas transaccionales, y una identificación clara de estas búsquedas a través de términos como "gratis, comprar, descargar y bajar".

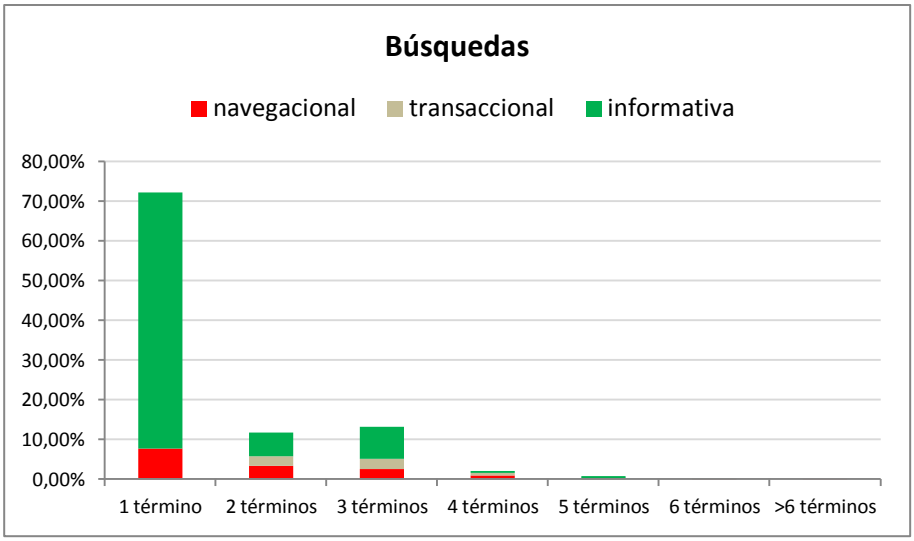


Figura 26. Intención de búsqueda por longitud de palabras clave

A medida que utilizamos más términos la estadística tiende a dar más protagonismo a la búsqueda navegacional, aunque en comparación, tanto el volumen de búsquedas como el número de palabras clave sea considerablemente menor que la búsqueda informativa.

6.1.4.1. BÚSQUEDA INFORMATIVA Y PALABRAS CLAVE

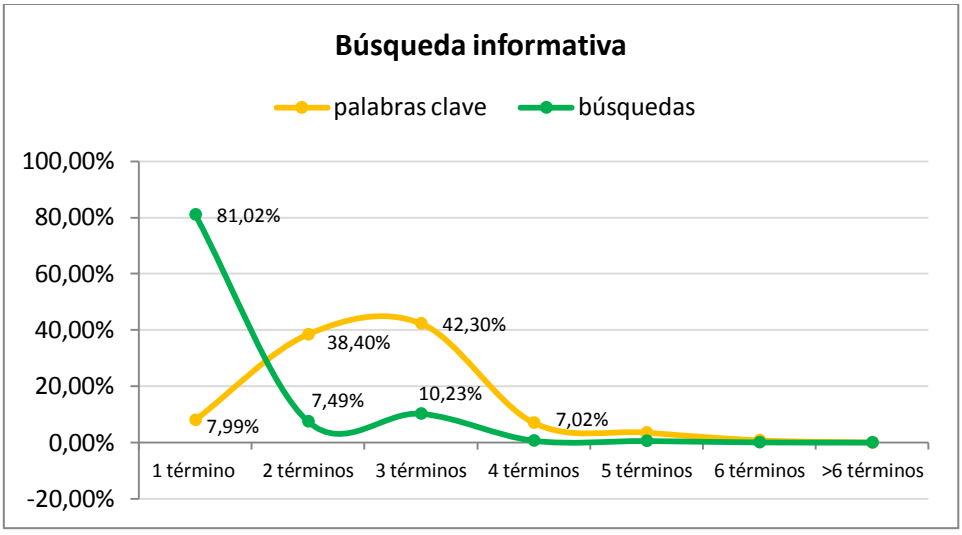


Figura 27. Palabras clave vs Búsquedas cuando se formulan búsquedas informativas

El 81,02 % de las búsquedas informativas se realizan utilizando cadenas de búsqueda de 1 solo término, y no hemos contado ninguna búsqueda informativa de más de 6 términos.

De las palabras clave utilizadas en las búsquedas informativas, el 80,70 % estaban formadas por 2 o 3 términos, y representan el 53,49 % del conjunto de palabras

clave. Más de la mitad de las palabras clave que estamos observando se han utilizado para formular el 14,10% de las búsquedas totales (datos de la tabla que se muestra en la Figura 26)

Parece fácil identificar las palabras clave de un solo término con las búsquedas informativas. Las navegacionales de 1 término son todas nombres propios o de instituciones, o hacen referencia a un dato concreto, por ejemplo editoriales como *mondadori*, *longman*, o autores como *coelho*, *bioy*,… o también *isbn*

6.1.4.2. BÚSQUEDA NAVEGACIONAL Y PALABRAS CLAVE

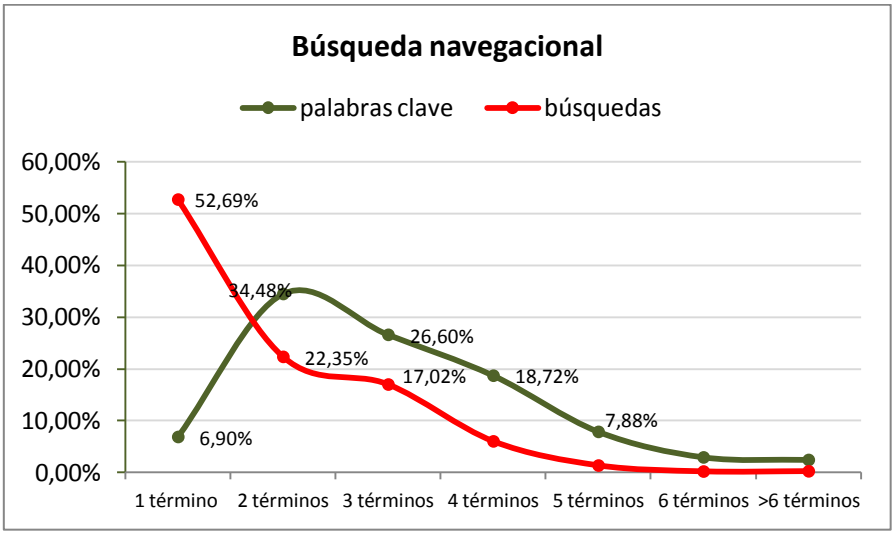


Figura 28. Palabras clave vs Búsquedas cuando se formulan búsquedas navegacionales

Aunque no es lo más habitual, una de las cosas más destacables es que las **búsquedas navegacionales** son las únicas que alguna vez utilizan más de 6 términos en la cadena de búsqueda.

Todas las búsquedas largas coinciden con títulos de libros:

la reina en el palacio de las corrientes de aire 10 términos

alicia en el pais de las maravillas libro 8 términos

un viejo que leia novelas de amor 7 términos

6.1.4.3. BÚSQUEDA TRANSACCIONAL Y PALABRAS CLAVE

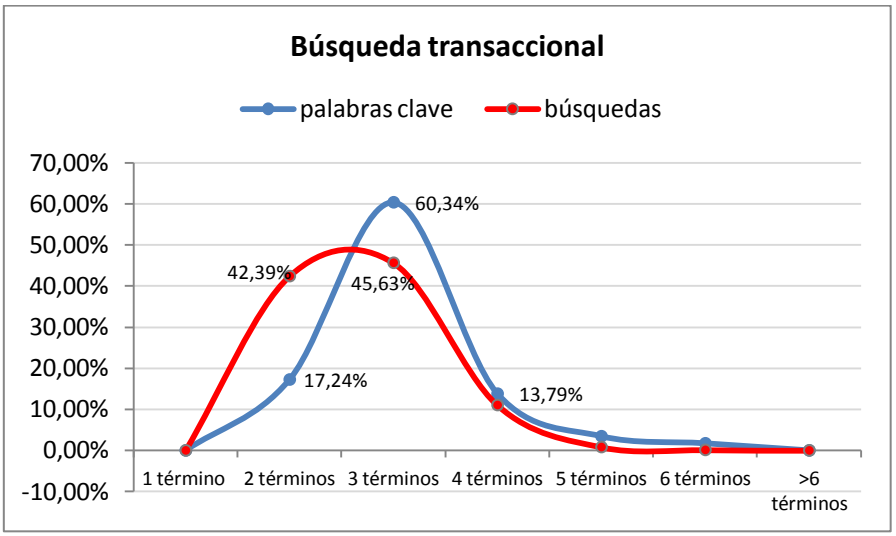


Figura 29. Palabras clave vs Búsquedas cuando se formulan búsquedas navegacionales

Para búsquedas transaccionales se utilizan palabras clave con una longitud mínima de 2 términos. No se pueden formular búsquedas transaccionales con 1 término. Necesitaremos al menos 1 término para describir la acción y otro para describir el contenido objeto de la acción.

Búsqueda transaccional= **acción + contenido**

6.1.5. PALABRAS CLAVE Y LEMBP

En este apartado se exponen los datos obtenidos al comparar las Palabras Clave que se han utilizado para realizar las búsquedas con los términos que se utilizan en las bibliotecas públicas para realizar la descripción de contenido, términos recogidos en la Lista de Encabezamientos de Materia de las Bibliotecas Públicas, LEMBP.

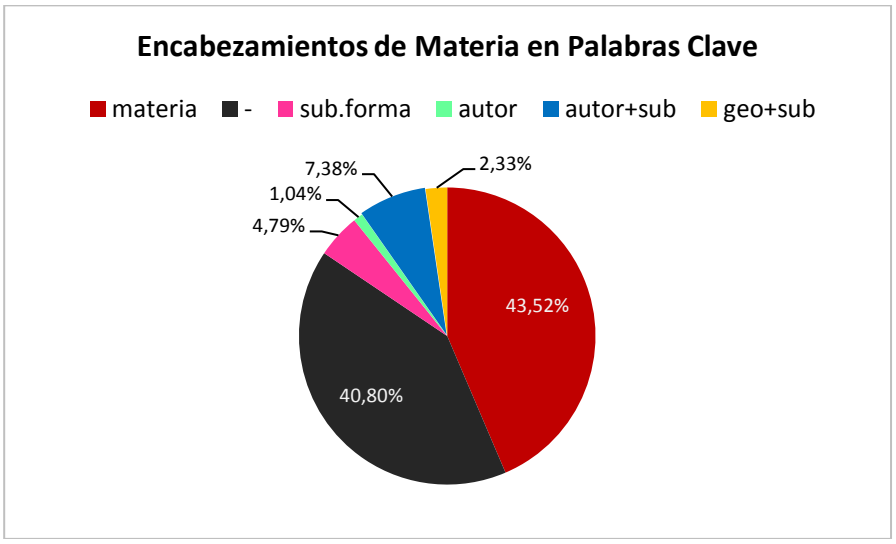


Figura 30. Palabras clave que utilizan términos recogidos en la LEMBP

El 40,80 % de las palabras clave no utiliza términos del LEMBP. En este grupo se han contabilizado las palabras clave que se utilizan con una intención de búsqueda transaccional, como *libros para descargar gratis* y aquellas otras que se utilizan con fines informativos o navegacionales, pero tienen construcciones como *aprender a leer* que describen acciones o deseos, u otras que hacen referencia al título de la obra como *libro crepúsculo*.

El resto, el otro **59,20 %** de las palabras clave analizadas utiliza, para la descripción de la búsqueda, algún término contemplado en la LEMBP:

El 43,52 % de las palabras clave son materias de la LEMBP:

premios nobel de literatura, *relato corto*, y también *tipos de lectura*.

[Lectura: Úsase para estudios sobre el arte de leer y la actitud del hombre frente a la lectura. LEMBP].

El 1,04 % de las palabras clave son nombres de autores. Sin subencabezamientos: *coelho*, *tatiana de*, *ursula k.leguin*

El 7,38 % de las palabras clave utilizan subencabezamientos para nombres de persona, y se han contabilizado tanto las palabras clave construidas con la estructura que marca la LEMBP como las que no:

nombre + subencabezamiento *garcia lorca obra*,

y también:

biografia del autor federico garcia lorca, , *libros de miguel Delibes*, *la poesia de lorca*,…

Se han agrupado las palabras clave que hacen referencia a Subencabezamientos bajo nombre de ciudades y Subencabezamientos bajo nombre de países en sub+geo; representan el 2,33 % de las Palabras Clave: *alejandria biblioteca*, *biblioteca santiago*, *biblioteca argentina*.

El 4,79 % de las palabras clave son materias que tienen también un subencabezamiento de forma:
antología poesía, *guía de lectura*, y también *libros digitales, audiolibros,* y *ereaders*

6.1.5.1. LOS SUBENCABEZAMIENTOS DE FORMA

Representan el 4,79 % del total de las palabras clave. Y en este grupo hemos incluido nuevos términos que no recoge la LEMBP.
Si estudiamos este grupo de forma aislada obtenemos los siguientes datos:

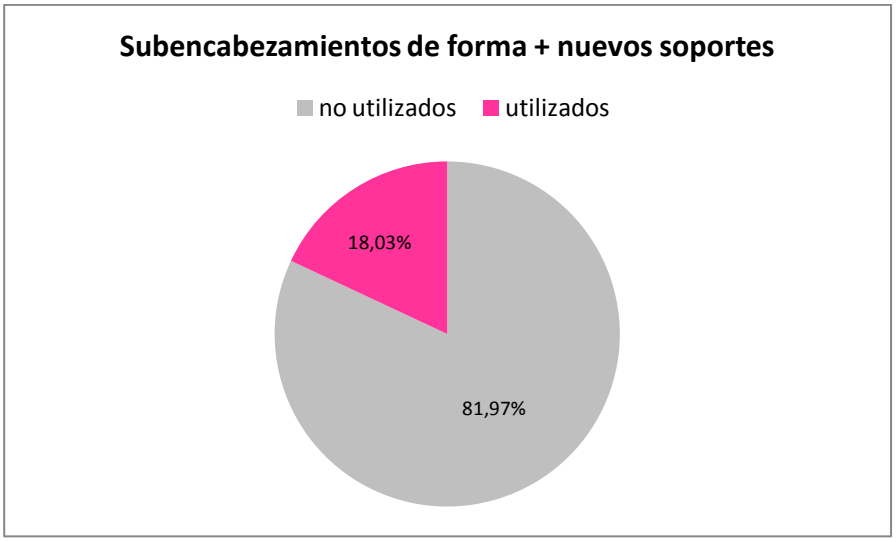


Figura 31. Palabras clave que utilizan algún término de la lista ampliada de subencabezamientos de forma

Solo hemos utilizado el 18,03 % de los términos de nuestra lista de subencabezamientos de forma, contando los términos *ebook, ereader, pdf, dvd, audiolibro, libros digitales, libros electrónicos, on line*.

Del listado original solo se utiliza el 13,16 % de los términos, que aglutinan la mayoría de los subencabezamientos utilizados: *novela*, *cuentos*, *ejercicios* y *certámentes*.

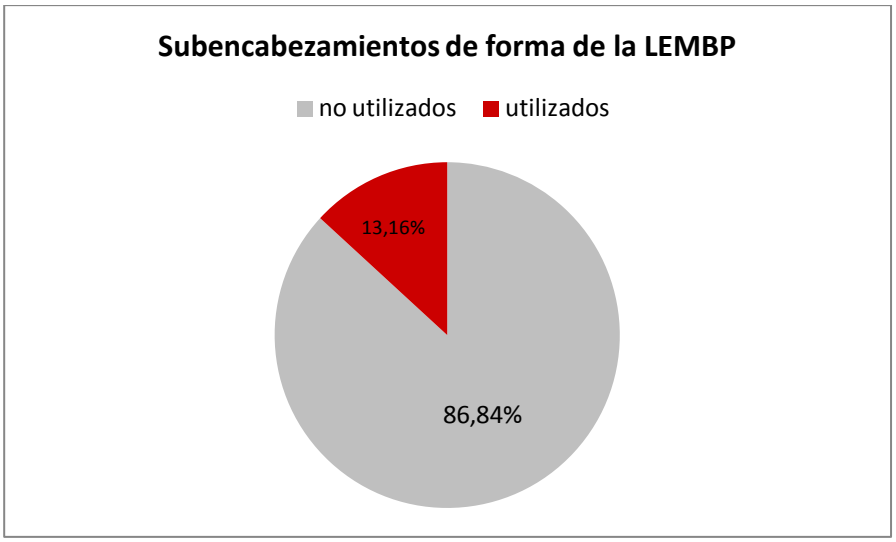


Figura 32. Palabras clave que utilizan algún término de la lista original de subencabezamientos de forma de la LEMBP

6.1.6. BÚSQUEDAS Y LEMBP

A continuación los datos referidos al volumen de búsquedas, es decir, cuántas veces se hace una búsqueda utilizando encabezamientos o subencabezamientos de materia.

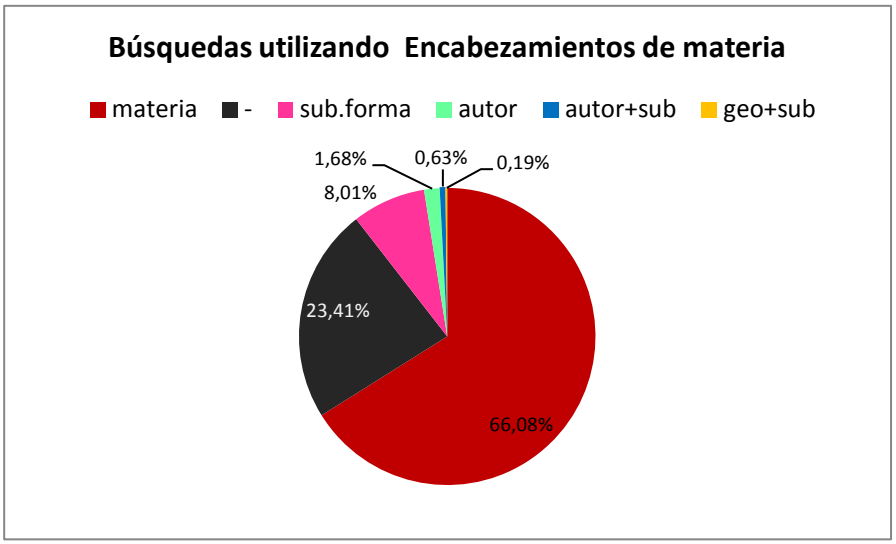


Figura 33. Búsquedas que utilizan términos recogidos en la LEMBP

El **76,59** % de las búsquedas se realiza utilizando algún término contemplado en la LEMBP.

Algunas formas de encabezamiento, como ciudad+subencabezamiento, país+subencabezamiento o autor, solo representan alrededor de un 2 % del total de búsquedas.

Pero el 74,09 % de las búsquedas contienen algún término recogido en la lista de encabezamientos de materia o subencabezamiento de forma.

Se han considerado materias las búsquedas con la siguiente estructura:

libros de ----- y *libros* ----, interpretando que *libros de* y *libros* contextualizan la búsqueda.

La LEMBP no recoge *libro* o *libros* como subencabezamiento de forma.

No es el caso de Novelas, que sí es un subencabezamiento de forma. Por eso, las búsquedas que emplean este término se han contabilizado como subencabezamientos de forma.

6.1.6.1. PRINCIPIOS BÁSICOS DE LA LEMBP

La LEMBP se basa en varios principios básicos, que delimitan el alcance de la lista y orientan al catalogador cuando se encuentra con distintas posibilidades de descripción del contenido de una obra. Esos principios son:

- Principio de especificidad: este principio se contempla desde 2 vertientes:
    - el elemento más específico por el que se puede definir el contenido es el nombre propio, que es el más específico y que no se contempla en la lista. Tampoco contiene "*nombres comunes* de cosas concretas enumerables en series uniformes", que serían propios de diccionarios enciclopédicos o manuales específicos de un tema, a los que se recomienda acudir, en vez de incluir los términos en la lista.
    - Cuando se utilice encabezamiento + subencabezamiento, en ningún caso se utilizará la sistematización en la descripción, es decir, el subencabezamiento no será una especificación de un término más genérico, el encabezamiento. El ejemplo que aparece en la LEMBP es "Puede estar justificado alguna vez, por tener pocos libros sobre este tema, catalogar un libro que trata de **Manzanos** bajo **Arboles frutales;** lo que nunca estará justificado es catalogarlo bajo **Arboles frutales-Manzanos.**"
- Principio de síntesis: "se intenta obtener la expresión del asunto o materia con un término de la máxima simplicidad".
- Principio de uso: "Los encabezamientos de materia deben acomodarse al fin para el que se utilizan: aplicarse a unos determinados fondos, ayudar a unos usuarios concretos, formar un catálogo llamado catálogo alfabético de materias." "... estos encabezamientos están pensados para macrodocumentos con predominio de los de tema genérico."
- Principio lingüístico: "Los encabezamientos se redactarán en español (cuando exista un término aceptado en nuestro idiomas), en el lenguaje usual (que no tiene que ser necesariamente el «vulgar») y en el orden natural del idioma (prevalencia del nombre sobre el adjetivo como elemento inicial)."

- Principio de uniformidad: “Cada materia debe tener siempre el mismo encabezamiento. Por consiguiente debe escogerse (basándose en razones de uso y de idioma) entre varios posibles (sinonimia) y debe distinguirse por medio de determinantes entre paréntesis en el caso de que un encabezamiento sirva para varias materias (homonimia).”
- Principio de economía: “Debe limitarse al máximo el número de encabezamientos para un mismo documento. Esto supone no acudir con asiduidad a los encabezamientos dobles o a los compuestos con «y» (que solucionan fácilmente el problema de las materias que se solapan), pero evitando al mismo tiempo el peligro de caer en los encabezamientos excesivamente genéricos.”

Se realiza una observación de las búsquedas desde el punto de vista del principio de síntesis, el principio de uso y el principio lingüístico, que son los que hacen referencia a la descripción propiamente dicha.

No tenemos un punto de partida para estudiar el principio de uniformidad, porque éste lleva implícita la existencia de una colección, y de unos bibliotecarios que catalogan con una uniformidad de criterio que no se puede dar cuando estudiamos el comportamiento de usuarios que actúan individualmente.

6.1.6.2. LAS BÚSQUEDAS Y EL PRINCIPIO DE ESPECIFICIDAD

Las búsquedas por nombre propio representan un 7,92% del total de búsquedas, e incluyen nombres propios de autores, lugares, comercios o títulos de obras. Como se explica en la LEMBP el nombre propio es el más específico de todos, aunque por razones obvias no se contemplen los nombres propios en la lista.

BÚSQUEDAS POR NOMBRE PROPIO	
Nombre propio de ...	Porcentaje de búsquedas
autor	4,91%
biblioteca	0,40%
editorial	0,69%
librería	0,28%
título	1,63%
TOTAL	7,92%

Figura 34. Porcentaje de búsquedas por nombre propio

Como se ha expuesto anteriormente, las palabras clave de 1 término de longitud, siendo el 7,12 % de las palabras clave del estudio acumulan el 72,18 % de las búsquedas. Pero si observamos este grupo, que es reducido en número de términos podemos observar que la mayoría son términos genéricos, opuestos a este principio de especificidad:

Palabra clave	Porcentaje de búsquedas
historia	14,56%
libro	7,95%
concursos	6,49%
ebook	6,49%
novela	4,37%
TOTAL	39,85%

Figura 35. Palabras clave más utilizadas

Hay 2 términos, *historia* y *libro*, que superan, cada uno de ellos, el volumen de búsquedas que se realizan por nombre propio. Con lo que podríamos decir que las búsquedas son más generales que específicas.

Podemos relacionar este resultado también con el principio de economía, que advierte del *peligro de caer en encabezamientos excesivamente genéricos* al intentar utilizar pocos términos en la descripción del contenido. Es lo que ocurre si utilizamos en exceso términos como *historia*, es una materia genérica, que nos da información de la materia, pero no despejará dudas sobre el tema tratado en el documento al que le asignemos éste encabezamiento.

6.1.6.3. LAS BÚSQUEDAS Y EL PRINCIPIO DE SÍNTESIS DE LA LEMBP

Principio de síntesis de la LEMBP y la longitud de los términos de búsqueda que utilizan términos de la Lista.

El 80,06 % de las búsquedas que utilizan algún encabezamiento de la LEMBP utiliza 1 solo término en la cadena de búsqueda.

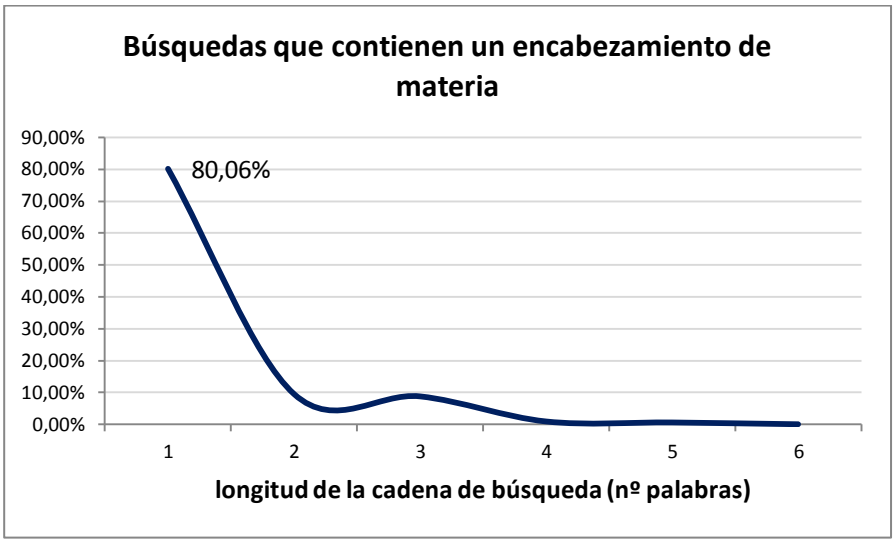


Figura 36. Principio de síntesis de las búsquedas que contienen términos del LEMBP

6.1.6.4. NUEVOS SOPORTES Y NUEVOS TÉRMINOS

El 8,07 % de las búsquedas hace referencia a nuevos soportes, no contemplados en la LEMBP. En la Figura 36 los porcentajes son sobre este 8,07 % de búsquedas.

En el caso de pdf, hace referencia a un formato de archivo, y en el siguiente gráfico, el 9 % que se le asigna representa la suma de las búsquedas asociadas a todas las palabras clave que hacen referencia al mismo: *libros pdf, libros en pdf, libros gratis pdf, libros en pdf gratis, libros gratis en pdf, libros pdf gratis, descargar libros pdf, bajar libros gratis en pdf, descargar libros gratis en pdf* y *libros gratis para descargar en pdf.*

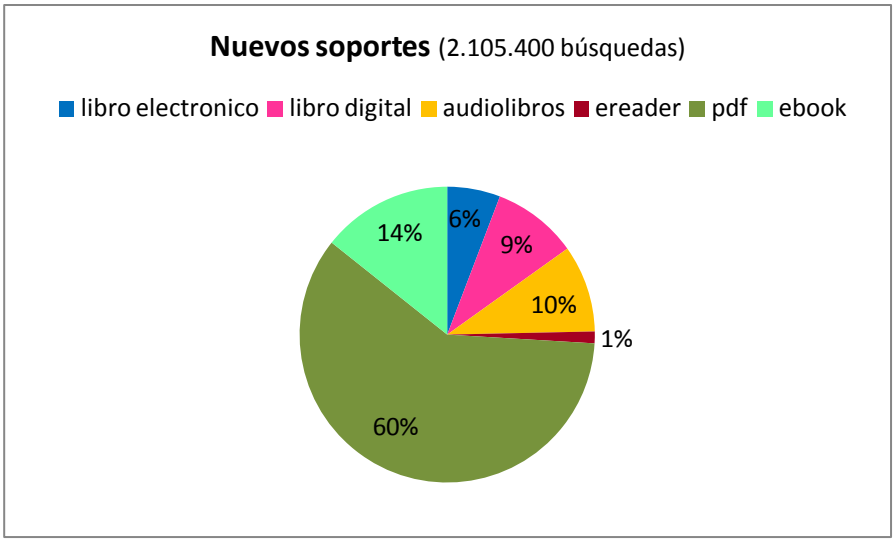


Figura 37. Nuevos términos incorporados a la lista de subencabezamientos de forma.

El término inglés *ebook* se utiliza indistintamente con *libro electrónico* y *libro digital*. Y el término *audiolibro*, que hace referencia a otro tipo de archivo, con un público distinto al del libro electrónico, representa el 10 % de las búsquedas.

6.1.6.5. PRINCIPIO LINGÜÍSTICO DE LA LEMBP Y EL IDIOMA DE LAS BÚSQUEDAS.

El 13,68 % de las búsquedas se realiza en inglés o combinando términos en inglés y español. Son palabras clave del tipo: *best seller books*, *library of congress*, *libros on line*, *spanish books*, *lector ebook*,…

6.1.7. OBSERVACIONES

Pero teniendo en cuenta que las búsquedas navegacionales representan el 14,72% de las búsquedas, ¿podemos concluir que los usuarios no "controlan" el tema de los libros? No. En este estudio hemos recontado un número de palabras clave que Adwords ha relacionado con un reducido grupo de palabras.

Por ejemplo, Adwords no ha relacionado *Harry Potter*, que tienen 823.000 de búsquedas locales mensuales con la palabra *bestseller*, ni con la palabra *libro*. Por tanto, *Harry Potter* no entra en nuestro estudio.

Del mismo modo tampoco aparece en nuestro listado de palabras clave *comic*, con 11.100.000 búsquedas globales mensuales y sí *ensayo*, con 1.830.000 búsquedas. Ambas son búsquedas informativas.

Lo que sí podemos afirmar es que hay un perfil de usuario que sí está más especializado. Si incluimos entre nuestras palabras clave el término cómic, y de ahí obtenemos sus palabras clave relacionadas observamos que el número de búsquedas navegacionales, el uso de nombres propios, nombres de libros y sagas es mucho mayor que entre las palabras clave que hemos utilizado finalmente en nuestro estudio.

¿Por qué eliminamos las palabras clave relacionadas con cómic? El del cómic es un segmento de mercado muy concreto, y distorsiona los resultados generales del estudio, ya que son fácilmente detectables los resultados relacionados con este género.

7. CONCLUSIONES

1. Hipótesis 1, no válida:

   "El usuario habitual de internet ha modificado su conducta, ya no busca igual que hace diez años."

   1.1. El usuario generalmente realiza búsquedas sencillas, casi siempre de un solo término. No encontramos grandes diferencias entre los resultados de estudios realizados hace diez años, referidos a búsquedas genéricas y éste, centrado en un tema específico.

   1.2. Han desaparecido completamente los operadores booleanos de las búsquedas, operadores con los que el usuario medio de la web no está nada familiarizado.

   1.3. El hecho de que la mayor parte de las búsquedas sean utilizando un solo término, y además, más de la mitad de estos términos sean encabezamientos de materia nos permite identificar grupos de interés. Hay ciertos temas que tienen prioridad respecto a otros: salud, aprendizaje infantil, los libros de autoayuda famosos,…

   No podemos dar como válida la primera hipótesis ya que no tenemos datos concluyentes sobre ellos. No tenemos datos anteriores sobre el mismo tema. Sí podemos decir cómo buscan los usuarios que buscan en español en Google.com sobre un tema determinado analizando una serie de términos utilizados, pero no podemos concluir que los hábitos de búsqueda hayan variado, a pesar de que algunos aspectos así nos lo indiquen.

2. Hipótesis 2, Válida:

   "El lenguaje utilizado por los usuarios que utilizan el castellano para realizar sus búsquedas en Google.com, sobre temática bibliotecaria, no coincide con los encabezamientos de materia de las bibliotecas públicas españolas, listado utilizado por los bibliotecarios de las bibliotecas públicas españolas y en castellano."

   2.1. Si hacemos una comparación literal de términos de búsqueda y encabezamientos de materia no podríamos confirmar la hipótesis 2 "el lenguaje utilizado en las búsquedas de usuario no coincide con los encabezamientos de materia".

Pero si hacemos ciertas consideraciones, como que el usuario no tiene por qué conocer las restricciones del uso de una lista de encabezamientos, la estructura del encabezamiento, y debemos asumir que el lenguaje que utiliza no está normalizado, sí podríamos dar por válida la hipótesis:

si observamos los principios de la Lista, vemos que efectivamente el usuario de internet sí cumple el principio de síntesis, ya que generalmente, el 80,06% de las veces, utiliza cadenas de búsqueda de 1 término, el principio lingüístico, ya que apenas un 13,68% de los términos se realizan en otro idioma que no sea el castellano, y la mayoría de estos términos son palabras que se han incorporado del inglés al castellano, y hacen referencia a nuevos soportes o formas de acceso a internet, como *online*, *ebook* o *best seller*.

Sólo en un 23,41 % de los casos no hace uso de términos o estructuras propias de la lista de encabezamientos.

2.2. Por otra parte es destacable que la lista de encabezamientos ofrece una amplia variedad de términos que no se utilizan en las búsquedas. Si observamos los subencabezamientos de forma veremos que solo se utilizan el 13,16 %. Y teniendo en cuenta que hemos asumido como subencabezamientos de forma algunos términos que hacen referencia a nuevos soportes que la Lista no recoge, interpretamos este dato como un síntoma de cambio, que hace necesaria una revisión de estos subencabezamientos, y no como desconocimiento por parte del usuario.

Algunos subencabezamientos entendemos que son de uso específico, *Cuadros sinópticos*, *Censos*, *Contestaciones para oposiciones*, y su uso viene determinado por un perfil de usuario más especializado, pero otros se han quedado directamente desfasados es el caso de *Videocassetes*, *Cintas magnetofónicas* o *Videodiscos*, que han sido reemplazados por *DVD's*, *PDF's*, o *audiolibros*.

2.3. El hecho de que las búsquedas informativas constituyan casi el 80 % del total de búsquedas encaja con la finalidad de la Lista de Encabezamientos, que es actuar como intermediario entre la biblioteca y el usuario para obtener la información necesaria sobre un tema. Se utiliza la cadena de búsqueda con la misma finalidad que los encabezamientos de materia, para identificar y llegar a información que satisfaga una necesidad de información.

Con los datos que hemos obtenido podemos concluir que el usuario no utiliza el mismo lenguaje que el propuesto en la LEMBP, pero sí comparte la idiosincrasia de la lista, como se ve en los datos sobre los principio de síntesis y lingüísticos y la incorporación de nuevos términos para la descripción de contenido.

7.1. PROPUESTA DE UNA METODOLOGÍA DE ANÁLISIS DE USUARIOS

Tras realizar este estudio proponemos una metodología para realizar análisis de usuarios de internet en base a lo que buscan y a lo que indizan a través del etiquetado, en las Webs 2.0.

Partiremos para esta propuesta de una serie de **premisas** que hemos manejado también para la realización de este estudio:

1. Para plantear un análisis de usuario previamente estableceremos los objetivos de dicho análisis, y de este modo dirigiremos nuestros esfuerzos hacia la respuesta que queramos encontrar.
2. Los objetivos que establezcamos determinarán los aspectos de las búsquedas a estudiar en cada caso.
3. No contamos con la información privilegiada que nos proporciona un fichero *log*. Si la tuviésemos, porque vamos a estudiar cómo buscan los usuarios en el buscador de nuestra página, por ejemplo, utilizaríamos los datos que nos proporciona el fichero como base de nuestro análisis y nos saltaríamos la primera parte, la de obtención de datos utilizando *Google Adwords*.
4. Las decisiones que se tomen con posterioridad son subjetivas y quedan fuera del alcance de este estudio y de la metodología de análisis propuesta.

7.1.1. Análisis de **usuario** en base a las búsquedas:

FASE 1. OBTENCIÓN DE DATOS

1. Definir objetivos. ¿qué queremos obtener con las búsquedas? Podemos estar trabajando en un proyecto para la mejora de nuestro sistema de recuperación de información, o para la adaptación de la estructura léxica de un catálogo, o de la navegación en un sitio web. Para diferentes objetivos estudiaremos distintos aspectos de las búsquedas.
2. Proceso de recopilación de datos a partir de la información que nos proporciona Google Adwords:
    a. Describir con varias palabras clave o expresiones el contenido de nuestra web o sistema de información.

b. Con los términos seleccionados, y utilizando de la herramienta de *Palabras Clave* de *Google Adwords*, obtener las expresiones de búsqueda con las que los usuarios de internet buscan contenido en teoría similar al que ofrecemos en nuestra web.

   En este paso tendremos que determinar el idioma, el ámbito geográfico de las búsquedas, filtraremos el valor mínimo de búsquedas globales o locales con el que queremos trabajar, si queremos incluir búsquedas que hagan referencia a contenido adulto y el tipo de dispositivo desde el que se realiza la búsqueda.

c. Exportar los resultados que vayamos obteniendo con cada término en formato *.csv, *Archivo de valores separados por comas* a Microsoft Office Excel.

d. Depurar los resultados obtenidos. Obtendremos las palabras clave definitivas unificando los listados obtenidos y eliminando duplicados y los campos de datos que no vayamos a utilizar.

FASE 2. ANÁLISIS Y EXTRACCIÓN DE INFORMACIÓN

3. Dirigir el análisis de datos hacia diferentes aspectos de las búsquedas en función de nuestras necesidades, que se definen antes de comenzar el análisis de las búsquedas. A partir de la información que obtenemos podemos estudiar:

   a. Vocabulario utilizado en las búsquedas

      En este caso queremos averiguar si el lenguaje que utilizamos nosotros en nuestro sistema de información es el mismo que utilizan nuestros usuarios en las búsquedas para describir los mismos ítems. Podemos estar hablando de un catálogo, tanto bibliográfico como comercial, o del vocabulario que utilizamos para nombrar las distintas secciones de nuestra web. En cualquier caso:

      i. generar un listado con el vocabulario que queremos testar, que puede ser:

         - lista de encabezamiento de materias
         - lista con el nombre de las secciones de un catálogo comercial

- lista con el nombre de los productos de un catálogo comercial
- lista con el nombre de las secciones, subsecciones, categorías, en definitiva una lista con el nombre de los distintos apartados de cualquier sistema jerárquico de organización de una web o un sistema de información.

ii. Comparar nuestro listado de nombres con el listado de búsquedas.

b. Intención de búsqueda

Establecer la intención de búsqueda que tienen nuestros usuarios potenciales al buscar contenido relacionado con el que ofrecemos en nuestra web. Nos da información sobre el tipo de información que busca el usuario, qué actividades están más relacionadas con nuestro contenido y también el conocimiento o la especialización del usuario.

c. Longitud de la cadena de búsqueda:

Contabilizar el número de palabras con las que nuestros usuarios potenciales realizan sus búsquedas.

Combinando los datos de longitud con los de intención de búsqueda obtenemos una visión más real de cómo busca el usuario. El esfuerzo de conceptualización o la relación entre longitud y especificidad de la búsqueda.

Además, podemos tomar decisiones de diseño del tipo *tamaño de la caja de búsqueda* o posición de nuestro buscador en la web si averiguamos el tipo de navegación que practicarán nuestros usuarios.

8. DISCUSIÓN

Sólo estamos estudiando búsquedas de “biblioteca”. No podemos extrapolar los datos a otros campos o temas de búsqueda, porque dependiendo de los términos que seleccionemos obtendremos una información u otra, como se ha visto con los resultados obtenidos estudiando la intención de búsqueda y que se han comparado con los resultados obtenidos en otros estudios.

Si queremos estudiar un grupo de usuarios interesados por ejemplo, en excursionismo tendremos que trabajar con un corpus de palabras relacionadas con ese tema. Del mismo modo es posible que otro estudio sobre bibliotecas escogiese otro corpus de palabras distinto del nuestro, con resultados distintos, y también sería válido. Las palabras que seleccionemos nos darán información sobre un aspecto concreto de un tema específico en el que queramos trabajar.

Respecto a la población estudiada, no hay error muestral ya que el 96,18 % de los usuarios que buscaron en español en internet durante el período observado, de enero a diciembre del 2010, utilizaron *Google* (StatCounter Global Stats, 2011), y por tanto el volumen de búsquedas para cada consulta o *query* estudiada es lo buscado por aquellos usuarios que buscan en Google en español esa consulta o *query* de entre el 96,18 % de usuarios que buscan (cualquier cosa) en español. *Porcentaje de usuarios de Google calculado en el Anexo 2.*

9. INVESTIGACIÓN FUTURA

Este estudio es una vista en conjunto de aspectos distintos que afectan a las búsquedas en internet. El reto ha sido realizarlo con datos de acceso abierto, datos a los que todos podemos acceder desde nuestra casa.

Estudios ampliados desde la especialización:

- Ampliar áreas geográficas de estudio, dentro de los usuarios de habla española, diferenciar usuarios españoles, mexicanos (que es la mayor comunidad hispanohablante), y del resto de países de habla española. Con la misma herramienta que hemos utilizado se puede especificar idioma y país desde el que se realizan las búsquedas.

- En estudios futuros será muy importante distinguir usuarios que acceden a internet desde un dispositivo fijo o uno móvil, encontrar diferencias en la formulación de la cadena de búsqueda y analizar si las necesidades de un mismo grupo de usuarios es distinta según el soporte desde el que se realiza la consulta.

- Análisis de búsquedas de temáticas concretas en el ámbito de las bibliotecas. Esta propuesta surge de resultados obtenidos a partir de la búsqueda de la palabra cómic, de la que se desprenden datos sustancialmente distintos a los obtenidos en este estudio (ver apartado 6.1.6), datos relacionados con la intención de búsqueda e incluso con la estructura de la cadena de búsqueda empleada para obtener información. La especialización del usuario provoca necesidades de información distintas, y por tanto, el modo de atender a esas necesidades de acceso a la información ha de ser distinto.

Estas tres propuestas se podrían llevar a cabo con datos de Google Adwords. Si quisiéramos llevar a cabo estos mismos estudios añadiendo variables relacionadas con el perfil de usuario, es decir, sexo, edad, nivel de estudios, sería necesario acceder a archivos log, que no son de acceso abierto.

Estudios paralelos:

- En la introducción hablamos de las actividades que realizan los usuarios en la web gracias a las herramientas 2.0 que tienen a su disposición, como el etiquetado social,

equiparable a la indización. Pero, el usuario convencional ¿etiqueta igual que el profesional? ¿El etiquetado nos ofrece más información que la indización, que utiliza vocabularios controlados? ¿Qué tipo de etiquetas se utilizan y cuáles pueden mejorar las búsquedas? porque si etiquetado social nos da la misma información que el profesional, ¿es necesario tenerlo en cuenta para mejorar nuestros sistemas de búsqueda?

BIBLIOGRAFÍA

- Baeza-Yates, R., Calderón-Benavides, L., González-Caro, C. (2006) The intention Behind Web Queries. En: *String Processing and Information Retrieval*, Vol. 4209, pp. 98-109. doi:10.1007/11880561_9
- Baeza-Yates, R. (2008). Web Mining or The Wisdom of the Crowds. In *Proceeding of the 2008 conference on Artificial Intelligence Research and Development: Proceedings of the 11th International Conference of the Catalan Association for Artificial Intelligence*, Teresa Alsinet, Josep Puyol-Gruart, and Carme Torras (Eds.). IOS Press, Amsterdam, The Netherlands, 3-3.
- Baeza-Yates, R. (2011a) Investigación en tecnologías de la información [vídeo] En: Jornadas Cátedra GedesTIC 2011-UPV Jornada 1. Sesión 1. Video 1/4. Disponible en: http://gedestic.es/videos/jornada-1-sesion-1-video-14-investigacion-en-tecnologias-de-la-informacion [consulta: 16/05/2011]
- Baeza-Yates, R. (2011a) Investigación en tecnologías de la información [vídeo] En: Jornadas Cátedra GedesTIC 2011-UPV Jornada 1. Sesión 1. Video 2/4. Disponible en: http://gedestic.es/videos/jornada-1-sesion-1-video-24-investigacion-en-tecnologias-de-la-informacion [consulta: 16/05/2011]
- Batelle, J. (2003) The Database of Intentions [en línea]. *John Batelle's Searchblog*. 13 noviembre 2003 http://battellemedia.com/archives/2003/11/the_database_of_intentions [consulta: 16/05/2011]
- Batelle, J. (2010) The Database of Intentions Is Far Larger Than I Thought [en línea]. *John Batelle's Searchblog*. 5 marzo 2010 http://battellemedia.com/archives/2010/03/the_database_of_intentions_is_far_larger_than_i_thought [consulta: 16/05/2011]
- Broder, A. (2002). A taxonomy of web search. *SIGIR Forum* 36, 2 (September 2002), Disponible en: http://www.sigir.org/forum/F2002/broder.pdf [consulta: 18/05/2011]
- Colaboradores de Wikipedia. (2011a)*Ley de potencias* [en línea]. Wikipedia, La enciclopedia libre, 2011. Disponible en: <http://es.wikipedia.org/w/index.php?title=Ley_de_potencias&oldid=46508039> [consulta: 20/05/2011]
- Colaboradores de Wikipedia.(2011b) *Motor de búsqueda* [Internet]. Wikipedia, La enciclopedia libre, 2011. Disponible en: http://es.wikipedia.org/w/index.php?title=Motor_de_b%C3%BAsqueda&oldid=46425393. [consulta: 18/05/2011].

- Colaboradores de Wikipedia, .(2011c) 'Distopía', *Wikipedia, La enciclopedia libre,* 15 abril 2011, 12:01 UTC, <http://es.wikipedia.org/w/index.php?title=Distop%C3%ADa&oldid=45640460> [descargado 6 junio 2011]
- DBLP Computer Science Bibliography, http://www.informatik.uni-trier.de/~ley/db/
- Dirección General del Libro y Bibliotecas (1994) *Lista de Encabezamientos de Materia para las Bibliotecas Públicas*. Disponible en: http://www.mcu.es/bibliotecas/MC/LEMBP/index.html
- Fundación Orange (2011) eespaña 2010, Informe anual sobre el desarrollo de sociedad de la información en España. Disponible en: http://www.informeeespana.es/docs/eE2010.pdf
- Golder, S. and Huberman, B. A. (2006). Usage Patterns of Collaborative Tagging Systems. Journal of Information Science, 32(2). 198-208. Disponible en: http://www.hpl.hp.com/research/idl/papers/tags/tags.pdf [Consultado 05/05/de 2010]
- Guinchat, C., Menou, M. (1983) Introducción general a las ciencias y técnicas de la información y de la documentación. Paris: UNESCO, 1983.
- Hassan Montero, Y. (2006a). Indización Social y Recuperación de Información. En: *No Solo Usabilidad*, nº 5, 2006. <nosolousabilidad.com>. ISSN 1886-8592 http://www.nosolousabilidad.com/articulos/indizacion_social.htm [consulta: 16/05/2011]
- Hassan Montero, Y. (2006b) Visualización y Recuperación de Información. En: *CTDI 2006*. ESIG, Portugal. 27 de Abril de 2006. http://www.nosolousabilidad.com/hassan/visualizacion_y_recuperacion_de_informacion.pdf [consulta: 16/05/2011]
- Huberman, B.A., Pirolli, P., Pitnow, J.E. and Lukose, R.M. (1998) Strong regularities in World Wide Web surfing. En: *Science, 280* (5360) 95-97.
- Jansen, B.J., Spink, A.(2006) How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing & Management. Volume 42, Issue 1, January 2006, Pages 248-263 doi:10.1016/j.ipm.2004.10.007
- Jansen, B. J., Booth, D. and Spink, A. (2007) Determining the user intent of web search engine queries. En: *Proceedings of the 16th international conference on World Wide Web* (WWW'07). ACM, New York, NY, USA, 1149-1150. DOI=10.1145/1242572.1242739 http://doi.acm.org/10.1145/1242572.1242739
- Jansen, J., Booth, D. (2010). Classifying web queries by topic and user intent. En: *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems* (CHI EA '10). ACM, New York, NY, USA, 4285-4290. DOI=10.1145/1753846.1754140 http://doi.acm.org/10.1145/1753846.1754140

- Jansen, J. (2011) Some folks who are mentioned in Understanding Sponsored Search [en línea]. *Web Search*. 14/05/2011 http://jimjansen.blogspot.com/2011/05/some-folks-who-are-mentioned-in.html [consulta 16/05/2011]
- Kroski, E. (2007) Folksonomies and user-based tagging. EN: N. Courtney ed. *Library 2.0 and beyond. Innovative Technologies and Tomorrow User*. Westport, Connecticut: Libraries Unlimited, pp. 91-103
- Lee, U., Liu, Z., Cho, J. (2005). Automatic identification of user goals in Web search. En: *Proceedings of the 14th international conference on World Wide Web* (WWW '05). ACM, New York, NY, USA, 391-400. http://doi.acm.org/10.1145/1060745.1060804
- LibraryThing. Disponible en http://www.librarything.com/zeitgeist
- Lozano, R. (2010) *Biblioteca, aprendizaje y acceso a la información en medio del temporal tecnológico*. En: Notas ThinkEpi 2010. Disponible en: http://www.thinkepi.net/biblioteca-aprendizaje-acceso-informacion-medio-temporal-tecnologico [Consultado 17/05/2011]
- Martín Fernández, F. J., Hassan Montero, Y. (2003). Conociendo a nuestros usuarios. En: *No Solo Usabilidad*, nº 2, 2003. <nosolousabilidad.com>. ISSN 1886-8592
- Ministerio de Cultura (2011) Definición de Biblioteca Pública. Disponible en: http://www.mcu.es/bibliotecas/MC/EBP/Definicion.html [consultado 02/06/2011]
- Morville, P., Rosenfeld, L. (2006) *Information architecture for the World Wide Web.* O'Reilly, 3ª edición.
- O'Reilly, T. (2006) *Qué es Web 2.0. Patrones del diseño y modelos del negocio para la siguiente generación del software*. Disponible en: http://sociedadinformacion.fundacion.telefonica.com/DYC/SHI/seccion=1188&idioma=es_ES&id=2009100116300061&activo=4.do?elem=2146
- Ronda León, Rodrigo; (2005). La Arquitectura de la Información y las Ciencias de la Información. En: *No Solo Usabilidad*, nº 4, 2005. ISSN 1886-8592. Disponible en: http://www.nosolousabilidad.com/articulos/ai_cc_informacion.htm [consultado 21/05/2011]
- Rovira, C., Codina, Ll., Marcos, M.C., Palma, M.V. (2004) *Información y documentación digital.* Barcelona: Documenta Universitaria
- Spink, A., Wolfram, D., Jansen, J. and Saracevic, T. (2001). Searching the Web: the public and their queries.En: *J. Am. Soc. Inf. Sci. Technol.* 52, 3 (February 2001), 226-234. DOI=10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.3.CO;2-I http://dx.doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.3.CO;2-

ANEXO 1

Bibliografía de los artículos citados en este trabajo, relacionados con las búsquedas web.

año 1945

- Bush, V. (1945). As we may think. Atlantic Monthly, 176(1), 101–108

año 1949

- Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley.

año 1967

- Glaser, B., & Strauss, A. (1967). The discovery of grounded theory: Strategies for qualitative research. Chicago, IL: Aldine Publishing Co

año 1969

- J.A. Swets, Effectiveness of information retrieval methods. American Documentation, 20, 72-89, 1969.

año 1972

- G. Salton, C.T. Yu and K. Lam, optimum term weighting. Journal of the ACM (forthcoming). K. Sparck Jones, A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 2_88, 11--21, 1972.

año 1975

- G. Salton, A Theory of indexing. Regional Conference Series in Applied Mathematics. Philadelphia: SIAM, 1975.
- Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 26(6), 321–343

año 1976

- S.E. Robertson, A theoretical model of the retrieval characteristics of information retrieval systems. Ph.D. Thesis, University of London, 1976.
- S.E. Robertson and K. Sparck Jones, Relevance weighting of 3earch terms. Journal of the Amezican Society for Information Science, 2_~7, 129- 146, 1976.

año 1977

- C.J. van Rijsbergen, A theoretical basis for the use of co-occurrence data in information retrieval. Journal of Documentation, 33, 106- 119, 1977.
- S.E. Robertson, Progress in documentation: Theories and models in information retrieval• Journal of Documentation, 33, 126-148, 1977.

año 1979

- van Rijsbergen, C. J. Information Retrieval. London: Butterworths, 1979
- W.B. Croft and D.J. Harper, Using probabilistic models of document retrieval without relevance information. Journal of Documentation, 35, 285- 2 79, 19 79.

año 1980

- G. Salton and H. Wu. Paper presented at the BCSACM Conference on Research and Development in Information Retrieval, Cambridge, June 1980. To appear in: Information Retrieval Research, Butterworths, forthcoming.

año 1981

- Sparck-Jones, K. (Ed.). (1981). Information Retrieval Experiments. London: Butterworth

año 1982

- Belkin, N., Oddy, R., & Brooks, H. (1982). ASK for information retrieval: Part I. background and theory. Journal of Documentation, 38(2), 61–71

año 1983

- Salton, G., & McGill, M. (1983). Introduction to Modern Information Retrieval. New York: McGraw-Hill

año 1988

- Saracevic, T., Kantor. P., Chamis, A., & Trivison, D. (1988). A study of information seeking and retrieving. I. Background and methodology Journal of the American Society for Information Science, 39(3), 161– 176

año 1989

- Nelson, M.J. Stochastic models for the distribution of index terms. *Journal of Documentation*, 45, 3 (1989) 227-237.
- Peters, T. (1989). When smart people fail: An analysis of the transaction log of an online public access catalog. Journal of Academic Librarianship, 15(6), 267–273

año 1992

- Harman, D. (1992). Overview of the second text retrieval conference (TREC-2). Retrieved 21 April 1999 from the World Wide Web from http://www.trec.nist.gov/pubs/tree2/papers/tx/01.txt
- Robertson, S.E., & Hancock-Beaulieu, M.M. (1992). On evaluation of IR systems. Information Processing and Management, 28(4), 457– 466
- Wolfram, D. (1992). Applying informetric characteristics of databases to IR system file design. Part I. Informetric models. *Information Processing and Management*, 28(1), 121-133.

año 1993

- Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. Journal of the American Society for Information Science, 44(3), 161–174
- Kaske, N. (1993). Research methodologies and transaction log analysis: Issues, questions, and a proposed model. Library Hi Tech, 11(2), 79–86
- Kurth, M. (1993). The limits & limitations of transaction log analysis Library Hi Tech, 11(2), 98–86
- Millsap, L., & Ferl, T. (1993). Search patterns of remote users: An analysis of OPAC transaction logs. Information Technology and Libraries, 11(3), 321–343
- Peters, T. (1993). The history & development of transaction log analysis Library Hi Tech, 42(11), 41–66
- J.R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
- Sandore, B. (1993). Applying the results of transaction log analysis Library Hi Tech, 11(2), 87–97
- Siegfried, S., Bates, M., & Wilde, D. (1993). A profile of end-user searching behavior by humanities scholars: The Getty online searching project report no. 2. Journal of the American Society for Information Science, 44(5), 273–291
- Wallace, P. (1993). How do patrons search the online catalog when no one's looking? Transaction log analysis and implications for bibliographic instruction and system design. RQ, 33(3), 239–252

año 1994

- M. K. Buckland and C. Plaunt. On the construction of Selection Systems. In *Library Hi Tech*, 12(4), 1994.

año 1995

- Catledge, L. D., & Pitkow, J. E. Characterizing Browsing Strategies in the World-Wide Web. *Proceedings of WWW3*, 1995.

- Croft, W., Cook, R., & Wilder, D. (1995). Providing government information on the Internet: Experiences with THOMAS. Paper presented at Digital Libraries Conference, Austin, TX
- Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via errorcorrecting output codes. Journal of AI Research 2 (1995) 263–286
- Saracevic, T. (1995). Evaluation of evaluation in information retrieval Paper presented at 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA
- Selberg, E., & Etzioni, O. (1995). Multi-service search and comparison using the Metacrawler. Proceedings of the Fourth World Wide Web Conference. Retrieved from the World Wide Web on 5 August 1999 from http://www.cern.ch/CERN/WorldWideWeb/Papers.html

año 1996

- Borgman, C. (1996). Why are online catalogs still hard to use? Journal of the American Society for Information Science, 47(7), 493–503
- Cockburn, A., & Jones, S. Which way now? Analysing and easing inadequacies in WWW navigation. *International Journal of Human-Computer Studies, 45*, 1996
- Crovella, M., & Bestavros, A. (1996). Self-similarity in World Wide Web traffic evidence and possible causes. Paper presented at ACM SIGMETRICS Conference on Measurement & Modeling of Computer Systems, Philadelphia, PA
- Hawkins, D. (1996). Hunting, grazing, browsing: A model for online information retrieval. Online, Retrieved from the World Wide Web on 9 July 1997 from http://www.onlineinc.com/onlinemag/
- Hoffman, D., Kalsbeek, W., & Novak, T. (1996). Internet and Web use in the U.S. Communications of the ACM, 39(12), 106–108
- Koenemann, J., & Belkin, N. (1996) A case for interaction: A study of interactive information retrieval behavior and effectiveness. Paper presented at conference on human factors in computing systems, Vancouver, Canada
- Saracevic, T. (1996). Modeling interaction in information retrieval: A review and proposal. Proceedings of the Annual Academy Meeting of American Society for Information Science (pp. 35–44)
- Zorn, P., Emanoil, M., & Marshall, L. (1996). Advanced searching: Tricks of the trade. Online. Retrieved from the World Wide Web on 1 July 1997 from http://www.onlineinc.com/onlinema/

año 1997

- Abdulla, G., Liu, B., Saad, R., & Fox, E. (1997). Characterizing World Wide Web queries. TR-97-04. Retrieved from the World Wide Web on 15 August 1999

- CommerceNet/Nielsen Media. (1997). Search engines most popular method of surfing the Web. Retrieved from the World Wide Web on 12 August 1999 from http://www.commerce.net/news/
- Keily, L. (1997). Improving resource discovery on the Internet: the user perspective. Proceedings of the 21st International Online Information Meeting (pp. 205–212)
- Korfhage, R. (1997). Information storage and retrieval. New York: Wiley
- Lesk, M. (1997). Going digital. Scientific American, 276(3), 58–60
- Lesk, M., Cutting, D., Pedersen, J., Noreault, T., & Koll, M. (1997). Panel Session on "real world" information retrieval. Panel presented at 20th Annual international ACM SIGIR conference on research and development in information retrieval. Philadelphia, PA
- Lynch, C. (1997). Searching the Internet. Scientific American, 276(3), 52–56
- Mizzaro, S. (1997). Relevance: The whole history. Journal of the American Society of Information Science, 48(9), 810–832
- Robertson, S. E., & Beaulieu, M. (1997). Research and evaluation in information retrieval. *Journal of Documentation*, *53*(1).
- Saracevic, T. & Kantor, P. (1997). Studying the value of library and information services. I. Establishing a theoretical framework. II. Methodology and Taxonomy. *Journal of the American Society for Information Science*, 48 (6), 527-542, 543-563.
- Sparck-Jones, K., & Willett, P. (Eds.). (1997). Readings in information retrieval. San Francisco: Morgan Kaufman
- Spink, A. & Saracevic, T. Interaction in information retrieval: Selection and effectiveness of search terms. *Journal of the American Society for Information Science*, 48, 8 (Aug. 1997) 728-740.
- B. Schneiderman, D. Byrd, and W. B. Croft. Clarifying search: A user-interface framework for text searches. *D-Lib Magazine*, January 1997.
- Tauscher, L., and Greenberg, S. How people revisit web pages: Empirical findings and implications for the design of history systems. *International Journal of Human-Computer Studies, 47*, 1997.

año 1998

- Abdulla, G., Liu, B., & Fox E. (1998). Searching the World-Wide Web: Implications from studying different user behavior. Paper presented at the World Conference of the World Wide Web, Internet, and Intranet, Orlando, FL
- S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the 7th International World Wide Web Conference*, 1998.
- Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. 2 (1998) 121–167

- Choo, C., Betlor, B., & Turnbull, D. (1998). A behavioral model of information seeking on the Web: Preliminary results of a study of how managers and IT specialists use the Web. Paper presented at the American Society of Information Science, Pittsburgh, PA
- Hoelscher, C. (1998). How Internet experts search for information on the Web. Paper presented at the World Conference of the World Wide Web, Internet, and Intranet, Orlando, FL
- Huberman, B.A., Pirolli, P., Pitnow, J.E. and Lukose, R.M. (1998) Strong regularities in World Wide Web surfing. *Science, 280* (5360) 95-97.
- Jansen, B.J., Spink, A., Bateman, J. and Saracevic, T. (1998). Real life information retrieval: A study of user queries on the Web. *SIGIR Forum*, 33 (1) 5-17.
- Jansen, B., Spink, A., Bateman, J., Sarevic, T. (1998) "Searchers, the subjects they search, and sufficiency: A Study of a large example of Excite searchers". World Conference on the WWW and Internet, Florida 1998.
- Jones, S., Cunningham, S., & McNab, R. (1998). An analysis of usage of a digital library. Proceeding of Second European Conference on Digital Libraries (pp. 261–277)
- Jones, S., Cunningham, S.J., and McNab, R. Usage analysis of a digital library. *Proceedings of the Third ACM Conference on Digital Libraries* (June 1998) 293-294.
- Kirsch, S. (1998). The future of Internet search (keynote address). Paper presented at the 21st Annual International ACM SIGIR conference on research and development in information retrieval, Melbourne, Australia Retrieved from the World Wide Web on 16 August 1999 from http://topgun.infoseek.com/stk/presentations/sigir.ppt
- Kirsch, S. "Infoseek's experiences searching the Internet". SIGIR FORUM Fall 98.
- J. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.
- Krishna, B., & Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. Proceedings of the 7th International World Wide Web Conference. Retrieved from the World Wide Web on 9 August 1999 from http://decweb.ethz.eh/WWW7/1937/
- Lawrence, S., & Giles, C. (1998). Searching the World Wide Web. Science 5360(28), 98–100
- Spink, A., Greisdorf, H., & Bateman, J. (1998). From highly relevant to not relevant: Examining different regions of relevance. Information Processing and Management, 34(5), 599–621
- Smith, T., Ruocco, A., & Jansen, B. (1998). Digital video in education Proceedings of the Thirtieth ACM SIGCSE Technical Symposium on Computer Science Education (pp. 122–126)
- Zumalt, J., & Pasicznyuk, R. (1998). The Internet and reference services: A real-world test of Internet utility. Reference and User Services Quarterly, 38(2), 165–172

año 1999

- Baeza-Yates, R., Ribeiro-Neto, B. *Modern Information Retrieval*, Ed. Addison Wesley, 1999.

- Byrne, M. D., John, B. E., Wehrle, N. S., and Crow, D. C. The tangled web we wove: A taskonomy of WWW use. In *Human Factors in Computing Systems: Proceedings of CHI 99* (pp. 544-551), Addison Wesley, 1999.
- Choo, C. W., Detlor, B., and Turnbull, D. . Information Seeking on the Web – An integrated model of browsing and searching. *Proceedings of the Annual Meeting of the American Society for Information Science (ASIS)*, 1999. Available at http://choo.fis.utoronto.ca/fis/respub/aisis99/
- Goodrum, A., & Spink, A. (1999). Visual information seeking: A study of image queries on the World Wide Web. Paper presented at the Annual Meeting of the American Society for Information Science, Washington, DC
- Gordon, M. and Pathak, P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing and Management, 35* (2) 141-180.
- Hawking, D., Craswell, N., Thistlewaite, P., Harman, D. (1999) *Results and challenges in web search evaluation*. WWW8, Toronto, pages 243-252. Elsevier
- Hofmann, T.: Probabilistic latent semantic analysis. In: Proc. of Uncertainty in Artificial Intelligence, Stockholm (1999)
- Jansen, B., Spink, A., & Saracevic, T. (1999). The use of relevance feedback on the Web: Implications for Web IR system design. Paper presented at the World Conference of the World Wide Web, Internet, and Intranet, October 24–30, 1999. Waikiki, Honolulu. Available at http://jimjansen.tripod.com/
- Jupiter Research. (1999). Go network announces new INFOSEEK Search: 30 percent faster, 50 percent larger. Retrieved from the World Wide Web on 5 September 2000 from http://info.go.com/press/search.html
- Kehoe, C., Pitkow, J., & Morton, K. (1999). Graphic, visualization, and usability center's 8th WWW user survey. Retrieved 15 August 1999 from the World Wide Web at http://www.gvu.gatech.edu/
- Lawrence, S., & Giles, C. (1999). Accessibility of information on the web Nature, 400, 107–109 Nielsen/NetRating. (1999). Retrieved from the World Wide Web on 11 August 1999 from http://www.nielsen-netratings.com/
- Leighton, H., Srivastava, J. (1999) *Precision amongWorld Wide Web Search Services (Search Engines)*. Journal of the American Society for Information Science, 50(10): 870-881, 1999.
- Navarro-Prieto, R., Scaife, M., & Rogers, Y. Cognitive Strategies in Web Searching. *Proceedings of the 5th Conference on Human Factors & the Web*, 1999. Available at http://zing.ncsl.nist.gov/hfweb/proceedings/navarro-prieto/index.html.
- NTIA. (1999). Defining the digital divide, the 3rd annual report by the National Telecommunications and Information Administration. Retrieved from the World Wide Web on 20 August 1999 from http:// www.ntia.doc.gov/ntiahome/digitaldivide/

- OCLC Inc. (1999). Web statistics and analysis. URL: http://www.oclc.org/oclc/research/projects/webstats/index.htm.
- Pharo, N. (1999). Solving problems on the World Wide Web. Internet Research: Electronic Networking Applications and Policy, 4(3). Retrieve from the World Wide Web on 1 August 1999 from http://www.shef.ac.uk/;is/publications/
- Piktow, J. (1999). Summary of WWW characteristics. The World Wide Web Journal. 2(2), 2–13
- Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large Web search engine query log. SIGIR Forum, 33(1), 6–12
- Spink, A., Bateman, J., & Jansen, B. (1999). Searching the Web: A survey of Excite users. Internet Research: Electronic Networking Applications and Policy. Retrieve from the World Wide Web on 7 August 1999 from http://www.shef.ac.uk/;is/publications/infres/ircont.html
- Spink, A., Bateman, J. & Jansen, B.J. (1999). Searching the Web: Survey of Excite users. *Internet Research: Electronic Networking Applications and Policies*, 9 (2) 117-128.
- Wolfram, D. (1999). Term co-occurrence in Internet search engine queries: An analysis of the Excite data set. Canadian Journal of Information and Library Science, 24(2/3), 12–33
- Xu, J.L. (1999). Internet search engines: Real world IR issues and challenges Paper presented at the Conference on Information and Knowledge Management. Kansas City, MO

año 2000

- Alexa Insider Page. (2000). Alexa Insider Side Bar. Retrieved from the World Wide Web on 30 March 2000 from http://insider.alexa.com/ insider?cli510
- D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In Proceedings of ACM SIGKDD '00, 2000.
- K. Bharat and A. Broder. Mirror, mirror, on the Web: A study of host pairs with replicated content. In Proceedings of the Eighth Int'l. World Wide Web Conf., 1999.
- N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines. Cambridge University Press, 2000.
- He, D., & Go¨ker, A. (2000). Detecting session boundaries from Web user logs. Paper presented at 22nd Annual colloquium of IR research, April 5– 7, 2000, Cambridge UK
- C. Holscher and G. Strube. Web search behaviour of Internet experts and Newbies. Proceedings of WWW9. 2000. Available at http://www9.org/w9cdrom/81/81.html.
- Jansen, B. J., Goodrum, A., Spink, A. (2000). Searching for multimedia: Video, audio, and image. Web queries. World Wide Web, 3(4)
- Jansen, B., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. Information Processing and Management, 36(2), 207–227
- Jansen, B. J., Spink, A., & Saracevic, T. (2000). A study of users queries on the Web. Information *Processing and Management*, 36 (2) (Special Issue: Web Research & IR).

- Ross, N., & Wolfram, D. (2000). End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine Journal of the American Society for Information Science, 51(10), 949– 958
- Spink A., Jansen, B.J. & Ozmultu, (2000). Query reformulation and relevance feedback by Excite users. *Internet Research: Electronic Networking Applications and Policies.*
- Statistical Research, Inc. (2000). New study shows Internet users are loyal to Web "Niches." Retrieved from the World Wide Web on 5 September 2000 from http://www.sriresearch.com/press/pr20000217.htm
- Sullivan, D. (2000). Search engine sizes. SearchEngineWatch.com. Retrieved from the World Wide Web on 29 April 2000 from http:// www.searchenginewatch.com/reports/sizes.html
- Xu, J.L. (2000). Multilingual search on the World Wide Web. Paper presented at the HICSS 33. Maui, HI

año 2001

- Amanda Spink, Dietmar Wolfram, Major B. J. Jansen, and Tefko Saracevic. 2001. Searching the Web: the public and their queries. *J. Am. Soc. Inf. Sci. Technol.* 52, 3 (February 2001), 226-234. DOI=10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.3.CO;2-I http://dx.doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.3.CO;2-I
- Bernard J. Jansen and Udo Pooch. 2001. A review of web searching studies and a framework for future research. *J. Am. Soc. Inf. Sci. Technol.* 52, 3 (February 2001), 235-246. DOI=10.1002/1097-4571(2000)9999:9999<::AID-ASI1607>3.3.CO;2-6 http://dx.doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1607>3.3.CO;2-6
- F. Cacheda, A. Viña. "Experiencies retrieving information in the World Wide Web", aceptado para ISCC 2001, 6th IEEE Symposium on Computers and Communications, 3-5 de Julio de 2001.
- J. Jackson. Pop Goes the Interstitial. *eMarketeer*, 7 June 2001, Available at http://www.emarketer.com/analysis/eadvertising/20010607_ead.html
- http://www.tic.udc.es/~fidel/docs/publications/jitel2001.pdf
- J. Muramatu and W. Pratt. Transparent queries: Investigating Users' Mental Models of Search Engines. Proceedings of SIGIR 2001.
- Nick Craswell, David Hawking and Stephen Robertson. Effective site finding using link anchor information. SIGIR'01. Available at http://pigfish.vic.cmis.csiro.au/~nickc//pubs/sigir01.pdf
- N. Craswell, D. Hawking and K. Griffiths. Which Search engine is best at finding airline site home pages?. CSIRO Mathematical and Information Sciences TR01/45, 2001. Available at http://pigfish.vic.cmis.csiro.au/~nickc//pubs/TR01-45.pdf
- D.Hawking, N. Craswell, and K. Griffiths. Which search engine is best at finding online services? WWW10 poster. Available at http://pigfish.vic.cmis.csiro.au/~nickc//pubs/www10actualposter.pdf

- Ponte, M., Cacheda, F., Viña, A. (2001) *Análisis de las búsquedas realizadas, categorías accedidas y documentos vistos en un directorio Web*.

año 2002

- Broder, A.: A taxonomy of web search. SIGIR Forum 36 (2002) 3–10
- A. Spink, B.J. Jansen, D. Wolfram, and T. Saracevic. From E-Sex to E-Commerce: Web search changes. IEEE Computer, 35(3):107 { 109, 2002.
- D.D. Wackerly, W. Mendenhall III, and R.L. Schea®er. Mathematical Statistics with Applications. Duxbury, 6th edition, 2002.

2003

- Basu, A., Watters, C., Shepherd, M.: Support vector machines for text categorization. In: International Conference on System Sciences, Washington, DC, USA, IEEE Computer Society (2003) 103.3
- Batelle, J. (2003) The Database of Intentions [en línea]. *John Batelle's Searchblog*. 13 noviembre 2003 http://battellemedia.com/archives/2003/11/the_database_of_intentions
- B.D. Davison, D.G. Deschenes, and D.B. Lewanda. Finding relevant Website queries. In Proceedings of the Twelfth Int'l. World Wide Web Conf., 2003.
- N. Eiron and K.S. McCurley. Analysis of anchor text for Web search. In Proceedings of ACM SIGIR '03, 2003.

año 2004

- Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query recommendation using query logs in search engines. In: Current Trends in Database Technology - EDBT, Springer-Verlag GmbH (2004) 588–596
- J.L. Devore. Probability and Statistics for Engineering and the Sciences. Duxbury, 6th edition, 2004.
- R. Kraft and J. Zien. Mining anchor text for query re¯nement. In Proceedings of the Thirteenth Int'l. World Wide Web Conf., 2004.
- Rose, D.E., Levinson, D.: Understanding user goals in web search. In: International conference on WWW, ACM Press (2004) 13–19
- Speretta, M., Gauch, S.: Personalizing search based on user search history (2004)
- Jin, X., Zhou, Y., Mobasher, B.: Web usage mining based on probabilistic latent semantic analysis. In: Knowledge discovery and data mining, New York, USA, ACM Press (2004) 197–205

año 2005

- Mobasher, B. In: Practical Handbook of Internet Computing. CRC Press (2005)

- Baeza-Yates, R.: Applications of web query mining. In: ECIR 2005. Volume 3408.,Lecture Notes in Computer Science (2005)
- Lee, U., Liu, Z., Cho, J.: Automatic identification of user goals in web search. In: International conference on WWW, ACM Press (2005) 391–400
- Lin, C., Xue, G.R., Zeng, H.J., Yu, Y.: Using probabilistic latent semantic analysis for personalized web search. In: Web Technologies Research and Development, Berlin Heidelberg, Springer-Verlag GmbH (2005) 707–717
- Jansen, B.J., Spink, A.: An analysis of web searching by european alltheweb.com users. Information Processing andManagement: an International Journal 41 (2005) 361–381

año 2006

- Baeza-Yates, R., Liliana Calderón-Benavides, L.,González-Caro, C. The intention Behind Web Queries. EN: *String Processing and Information Retrieval* In String Processing and Information Retrieval, Vol. 4209 (2006), pp. 98-109. doi:10.1007/11880561_9 Key: citeulike:2435010
- Jansen, B.J., Spink, A.(2006) How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing & Management. Volume 42, Issue 1, January 2006, Pages 248-263 doi:10.1016/j.ipm.2004.10.007
- O'Reilly, T. (2006) *Qué es Web 2.0. Patrones del diseño y modelos del negocio para la siguiente generación del software*. Disponible en: http://sociedadinformacion.fundacion.telefonica.com/DYC/SHI/seccion=1188&idioma=es_ES&id=2009100116300061&activo=4.do?elem=2146
- Shen, D., Pan, R., Sun, J.-T., Pan, J. J., Wu, K., Yin, J. and Yang, Q. Query enrichment for web-query classification *Transactions on Information Systems*, 24, 3 (2006), 320 - 352.

año 2007

- Beitzel, S. M., Jensen, E. C., Lewis, D. D., Chowdhury, A. and Frieder, O. Automatic classification of Web queries using very large unlabeled query logs *ACM Transactions on Information Systems*, 25, 2 (2007), Article No. 9.
- Beitzel, S. M., Jensen, E. C., Chowdhury, A. and Frieder, O. Varying approaches to topical web query classification. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (Amsterdam, The Netherlands, 23 - 27 July, 2007), 783 - 784.
- Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. 2007. Determining the user intent of web search engine queries. In *Proceedings of the 16th international conference on World Wide Web* (WWW '07). ACM, New York, NY, USA, 1149-1150. DOI=10.1145/1242572.1242739 http://doi.acm.org/10.1145/1242572.12427392009
- Liu, J., Birnbaum, L. (2007) Measuring Semantic Similarity between Named Entities by Searching the Web Directory. En: 2007 IEEE / WIC / ACM International Conference on Web Intelligence,

WI 2007, 2-5 November 2007, Silicon Valley, CA, USA, Main Conference Proceedings 2007 doi: 10.1109/WI.2007.75

- Meredith Ringel Morris and Eric Horvitz. 2007. SearchTogether: an interface for collaborative web search. In *Proceedings of the 20th annual ACM symposium on User interface software and technology* (UIST '07). ACM, New York, NY, USA, 3-12. DOI=10.1145/1294211.1294215 http://doi.acm.org/10.1145/1294211.1294215
- Bamba, B., Ling Liu, Caverlee, J., Padliya, V., Srivatsa, M., Bansal, T., Palekar, M., Patrao, J., Suiyang Li, Singh, A. (2007) *DSphere: A Source-Centric Approach to Crawling, Indexing and Searching the World Wide Web.* doi: db/conf/icde/icde2007.html#BambaLCPSBPPLS07
- Kroski, E. (2007) Folksonomies and user-based tagging. EN: N. Courtney ed. *Library 2.0 and beyond. Innovative Technologies and Tomorrow User*. Westport, Connecticut: Libraries Unlimited, pp. 91-103
- Pirolli, P. *Information Foraging Theory: Adaptive Interaction with Information*. Oxford University Press, Oxford, 2007.

año 2008

- Xiao-Bing Xue, Zhi-Hua Zhou, and Zhongfei (Mark) Zhang. 2008. Improving Web search using image snippets. *ACM Trans. Internet Technol.* 8, 4, Article 21 (October 2008), 28 pages. DOI=10.1145/1391949.1391955 http://doi.acm.org/10.1145/1391949.1391955
- Costa, R.P., Seco, N.(2008) Hyponymy Extraction and Web Search Behavior Analysis Based on Query Reformulation. EN: Advances in Artificial Intelligence – IBERAMIA 2008 Lecture Notes in Computer Science, Volume 5290/2008, 332-341, DOI: 10.1007/978-3-540-88309-8_34
- Caviglia, F., Ferraris, M. (2008) The Web as a learning environment: Focus on contents vs. focus on the search process EN: *Learning to Live in the Knowledge Society* IFIP International Federation for Information Processing, Volume 281/2008, 175-178, DOI: 10.1007/978-0-387-09729-9_27
- Baeza-Yates, R. (2008). Web Mining or The Wisdom of the Crowds. In *Proceeding of the 2008 conference on Artificial Intelligence Research and Development: Proceedings of the 11th International Conference of the Catalan Association for Artificial Intelligence*, Teresa Alsinet, Josep Puyol-Gruart, and Carme Torras (Eds.). IOS Press, Amsterdam, The Netherlands, The Netherlands, 3-3.

año 2009

- Duda, C., Frey, G., Kossmann, D., Matter, R., Chong Zhou (2009) AJAX Crawl: Making AJAX Applications Searchable EN: Data Engineering, 2009. ICDE '09. IEEE 25th International Conference on: 78 – 89. doi: 10.1109/ICDE.2009.90

- Lozano, R. (2010) *Biblioteca, aprendizaje y acceso a la información en medio del temporal tecnológico*. En: Notas ThinkEpi 2010. Disponible en: http://www.thinkepi.net/biblioteca-aprendizaje-acceso-informacion-medio-temporal-tecnologico

año 2010

- Chowdhury, G. (2010), Social Information Retrieval Systems: Emerging Technologies and Applications for Searching the Web Effectively. Journal of the American Society for Information Science and Technology, 61: 2587–2588. doi: 10.1002/asi.21410
- Baeza-Yates, R. (2010). *Searching the Web of Objects.* EN: Objects and Databases Lecture Notes in Computer Science, 2010, Volume 6348/2010, 6-7, DOI: 10.1007/978-3-642-16092-9_2
- Takata, N., Ohshima, H., Oyama, S., Tanaka, K. (2010) Searching the Web for Alternative Answers to Questions on WebQA Sites EN: Web-Age Information Management Lecture Notes in Computer Science, 2010, Volume 6184/2010, 441-452, DOI: 10.1007/978-3-642-14246-8_43
- Ioannidis, S., Massoulie, L. (2010) Surfing the Blogosphere: Optimal Personalized Strategies for Searching the Web EN: INFOCOM, 2010 Proceedings IEEE, 1 – 9, DOI: 10.1109/INFCOM.2010.5462079
- Jae-wook Ahn and Peter Brusilovsky. 2010. What you see is what you search: adaptive visual search framework for the web. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 1049-1050. DOI=10.1145/1772690.1772798 http://doi.acm.org/10.1145/1772690.1772798
- Rob Ennals, Dan Byler, John Mark Agosta, and Barbara Rosario. 2010. What is disputed on the web?. In *Proceedings of the 4th workshop on Information credibility* (WICOW '10). ACM, New York, NY, USA, 67-74. DOI=10.1145/1772938.1772952 http://doi.acm.org/10.1145/1772938.1772952
- Jansen, J., Booth, D. (2010). Classifying web queries by topic and user intent. En: *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems* (CHI EA '10). ACM, New York, NY, USA, 4285-4290. DOI=10.1145/1753846.1754140 http://doi.acm.org/10.1145/1753846.1754140

año 2011

- Jansen, J. (2011) Some folks who are mentioned in Understanding Sponsored Search [en línea]. *Web Search*. 14 mayo 2011 http://jimjansen.blogspot.com/2011/05/some-folks-who-are-mentioned-in.html [consulta 16 mayo 2011]
- Baeza-Yates, R. (2011a) Investigación en tecnologías de la información [vídeo] En: Jornadas Cátedra GedesTIC 2011-UPV Jornada 1. Sesión 1. Video 1/4. Disponible en: http://gedestic.es/videos/jornada-1-sesion-1-video-14-investigacion-en-tecnologias-de-la-informacion [consulta: 16 mayo 2011]

- Baeza-Yates, R. (2011a) Investigación en tecnologías de la información [vídeo] En: Jornadas Cátedra GedesTIC 2011-UPV Jornada 1. Sesión 1. Video 2/4. Disponible en: http://gedestic.es/videos/jornada-1-sesion-1-video-24-investigacion-en-tecnologias-de-la-informacion [consulta: 16 mayo 2011]

ANEXO 2

Cálculo del porcentaje de población mundial que utiliza el español como idioma en Internet y que además utiliza Google como buscador:

Para obtener este dato consultaremos la web http://www.internetworldstats.com, para obtener el número de países que utilizan el español como lengua de navegación por internet.

Una vez obtengamos los datos de los países y el porcentaje que representa cada uno, consultamos país por país, cuál es el buscador más utilizado en el período de tiempo coincidente con la recogida de datos de este estudio, entre enero y diciembre de 2010. Esta información la obtenemos en http://gs.statcounter.com

No tenemos datos globales sobre el uso de buscadores, por eso tenemos que hacer la búsqueda por países.

Por último, obtenemos el promedio de uso de Google en cada país, y ese dato lo ponderaremos de acuerdo con la población que representa cada uno de ellos. De este modo obtendremos el dato de la población con la que estamos trabajando.

Países que utilizan el español como idioma de Internet:

USUARIOS DE INTERNET HISPANOHABLANTES					
COUNTRIES	Population (2009 Est.)	Internet Users, Latest Data	Penetration % Population	User Growth 2000-2009	Table % Users
Argentina	40,913,584	20,000,000	48.9 %	700.0 %	14.3 %
Bolivia	9,775,246	1,050,000	10.7 %	775.0 %	0.8 %
Chile	16,601,707	8,369,036	50.4 %	376.2 %	6.0 %
Colombia	43,677,372	20,788,818	47.6 %	2,267.7 %	14.9 %
Costa Rica	4,253,877	1,500,000	35.3 %	500.0 %	1.1 %
Cuba	11,451,652	1,450,000	12.7 %	2,316.7 %	1.0 %
Dominican Republic	9,650,054	3,000,000	31.1 %	5,354.5 %	2.1 %
Ecuador	14,573,101	1,840,678	12.6 %	922.6 %	1.3 %
El Salvador	7,185,218	975	13.6 %	2,337.5 %	0.7 %
Equatorial Guinea	633,441	12	1.9 %	2,300.0 %	0.0 %
Guatemala	13,276,517	1,960,000	14.8 %	2,915.4 %	1.4 %
Honduras	7,833,696	958,5	12.2 %	2,296,3 %	0.7 %
Mexico	111,211,789	27,600,000	24.8 %	917.5 %	19.7 %
Nicaragua	5,891,199	600	10.2 %	1,100.0 %	0.4 %
Panama	3,360,474	934,5	27.8 %	1,976.7 %	0.7 %
Paraguay	6,995,655	894,2	12.8 %	4,371.0 %	0.6 %
Perú	u29,546,963	7,636,400	25.8 %	205.5 %	5.5 %
Puerto Rico	3,966,213	1,000,000	25.2 %	400.0 %	0.7 %
Spain	40,525,002	29,093,984	71.8 %	440.0 %	20.8 %
Uruguay	3,494,382	1,340,000	38.3 %	262.2 %	1.0 %
Venezuela	26,814,843	8,846,535	33.0 %	831.2 %	6.3 %
TOTAL	411,631,985	139,849,651	34.0 %	669.2 %	100.0 %

Figura 38. Usuarios de internet hispanohablantes

Fuente:

A continuación, los datos del porcentaje de uso de cada buscador en el período enero 2010-diciembre 2010 en los países de la tabla anterior:

	BUSCADOR					
País	**Google**	**bing**	**Yahoo!**	**Ask Jeeves**	**AltaVista**	**Other**
España	97,24 %	1,48 %	0,83 %	0,39 %	0,01 %	0,04 %
Argentina	96,39 %	1,72 %	1,59 %	0,21 %	0,04 %	0,03 %
Bolivia	97,16 %	2,06 %	0,4767	0,23 %	0,03 %	0,03 %
Chile	98,31 %	1,15 %	0,2542	0,24 %	0,02 %	0,02 %
Colombia	96,77 %	2,36 %	0,4925	0,30 %	0,03 %	0,02 %
Costa rica	95,32 %	2,28 %	1,9017	0,38 %	0,06 %	0,05 %
Cuba	96,55 %	1,90 %	1,16	0,23 %	0,07 %	0,08 %
Ecuador	96,07 %	2,48 %	0,9292	0,44 %	0,03 %	0,03 %
El salvador	94,98 %	2,71 %	1,8442	0,40 %	0,03 %	0,03 %
Guatemala	94,77 %	3,09 %	1,66	0,39 %	0,04 %	0,03 %
Honduras	96,35 %	1,11 %	0,3258	0,04 %	0,04 %	0,06 %
Guinea	85,34 %	6,30 %	6,2067	0,96 %	0,94 %	0,24 %
Mexico	93,36 %	4,28 %	2,0475	0,22 %	0,05 %	0,03 %
Nicaragua	92,80 %	3,63 %	2,95	0,53 %	0,03 %	0,05 %
Panama	95,09 %	2,85 %	1,515	0,37 %	0,10 %	0,06 %
Paraguay	96,90 %	2,12 %	0,5642	0,38 %	0,01 %	0,02 %
Perú	98,39 %	1,11 %	0,2892	0,16 %	0,02 %	0,02 %
Puerto rico	85,76 %	8,20 %	5,3292	0,46 %	0,13 %	0,11 %
Rep. Dominicana	95,51 %	2,91 %	1,0525	0,44 %	0,03 %	0,05 %
Uruguay	97,46 %	1,64 %	0,54	0,31 %	0,02 %	0,03 %
Venezuela	97,44 %	1,73 %	0,4658	0,31 %	0,01 %	0,03 %

Figura 39. Buscadores más usados por países hispanohablantes en 2010

Utilizamos el dato de porcentaje de usuarios de la tabla “Usuarios De Internet Hispanohablantes” para ponderar el dato de porcentaje de uso de Google de la tabla de uso de buscadores. De este modo sacamos el porcentaje de uso de Google entre los usuarios hispanohablantes de internet.

Porcentaje de uso de Google en español: 96,18%

País	Google	Valor ponderación	% total
España	97,24 %	14.3 %	20,23
Argentina	96,39 %	0.8 %	13,78
Bolivia	97,16 %	6.0 %	0,78
Chile	98,31 %	14.9 %	5,90
Colombia	96,77 %	1.1 %	14,42
Costa rica	95,32 %	1.0 %	1,05
Cuba	96,55 %	2.1 %	0,97
Ecuador	96,07 %	1.3 %	1,25
El salvador	94,98 %	0.7 %	0,66
Guatemala	94,77 %	0.0 %	1,33
Honduras	96,35 %	1.4 %	0,67
Guinea	85,34 %	0.7 %	0,00
Mexico	93,36 %	19.7 %	18,39
Nicaragua	92,80 %	0.4 %	0,37
Panama	95,09 %	0.7 %	0,67
Paraguay	96,90 %	0.6 %	0,58
Perú	98,39 %	5.5 %	5,41
Puerto rico	85,76 %	0.7 %	0,60
Rep. Dominicana	95,51 %	20.8 %	2,01
Uruguay	97,46 %	1.0 %	0,97
Venezuela	97,44 %	6.3 %	6,14
		TOTAL	96,18%

Figura 40. Tabla ponderada uso de Google en español