

# Análisis de Consultas a un Buscador y su Aplicación a la Jerarquización de Páginas Web\*

Ricardo Baeza-Yates      Felipe Saint-Jean

Depto. de Ciencias de la Computación  
Universidad de Chile  
Blanco Encalada 2120, Santiago, Chile  
{rbaeza,fsaint}@dcc.uchile.cl

## Resumen

Existen diversas fuentes de información en la Web. Una de ellas es la información que deja el usuario mientras usa un buscador. Es posible utilizar esta información para complementar el resultado de algoritmos de jerarquización tradicionales, de manera de agregar conocimiento humano al resultado. Dentro de este concepto, en el presente trabajo se muestra una visión global de los análisis de bitácoras de acceso al buscador TodoCl, mostrando de forma general qué es lo que buscan los usuarios en la Web Chilena. Luego se expone una aplicación práctica de los datos, al ver los resultados de la implementación de un algoritmo que determina relevancia y jerarquía de documentos basándose en las búsquedas y documentos seleccionados por los usuarios.

Palabras claves: buscadores Web, análisis de *logs*, jerarquización de páginas Web.

## 1 Introducción

Del amplio espectro de datos existentes en la Web que pueden ser analizados para extraer información, los datos de uso de la Web son de los más interesantes pues representan los intereses actuales de la comunidad de usuarios de un sitio. Existen numerosos trabajos sobre análisis de datos de navegación en un sitio y su uso para el rediseño del mismo, pero no sobre uso de servicios específicos, como por ejemplo, buscadores de la Web (una excepción es [1]).

En este estudio se detalla el análisis realizado de las acciones del usuario al utilizar un buscador, en nuestro caso un buscador de la Web chilena: TodoCL [4]. Esto es interesante desde dos puntos de vista. Por una parte mejora la calidad del sitio dando una mejor experiencia de búsqueda al usuario del buscador. Por otra parte, el usuario tiene alta comprensión semántica de los recursos y consultas involucradas, por lo que su navegación y utilización deja conocimiento que puede ser utilizado en los procesos de búsqueda y jerarquización de documentos.

Para nuestro análisis usamos 777.351 consultas realizadas en agosto y septiembre del 2001, las cuales contenían 738.390 palabras, con un vocabulario único de 465.021 palabras. Es decir, en promedio, sólo se usan 1.05 palabras por consulta. En la bitácora (*log*) de consultas, el buscador también registra la navegación que se realiza, en particular las páginas escogidas por los usuarios. La mayoría de las personas refina su consulta agregando o eliminando palabras, pero en promedio sólo inspecciona 1.15 páginas de resultados. Estudios similares han sido realizados con datos de Altavista [2] y Excite [5, 3]. En nuestro caso agregamos un análisis de variabilidad de las consultas y también usamos nuestros resultados para mejorar la jerarquización (*ranking*) de páginas, experimentando con un algoritmo recientemente propuesto [6].

En la siguiente sección presentamos el análisis de consultas, incluyendo palabras más buscadas, su distribución y variabilidad, opciones de consulta y la relación de las consultas y las páginas en la Web chilena.

---

\* Parcialmente financiado por Proyecto Fondecyt 1020803, Chile.

En la sección 3 usamos un algoritmo que permite mejorar el *ranking* de páginas relacionadas con la consulta usando las páginas escogidas por los usuarios. En la sección final entregamos nuestras conclusiones y extensiones futuras.

## 2 Navegación en el Buscador

Los usuarios, al acceder al sitio, dejan su huella en las bitácoras de acceso. A partir de ellos es posible reconstruir el camino seguido por cada usuario. La reconstrucción no es tan directa como aparenta, ya que el *cache* del navegador hace que ciertos pasos del usuario sean omitidos de las bitácoras. Por ejemplo, es posible que el usuario haya escrito la URL directamente en su browser (no siguiendo un enlace) o el cache del navegador cargó la página sin pedirla al servidor. En general, se consideró que el caso en que el usuario ingresa la URL directamente es menor, por lo que la mayoría de los pasos de navegación siguen la estructura del sitio.

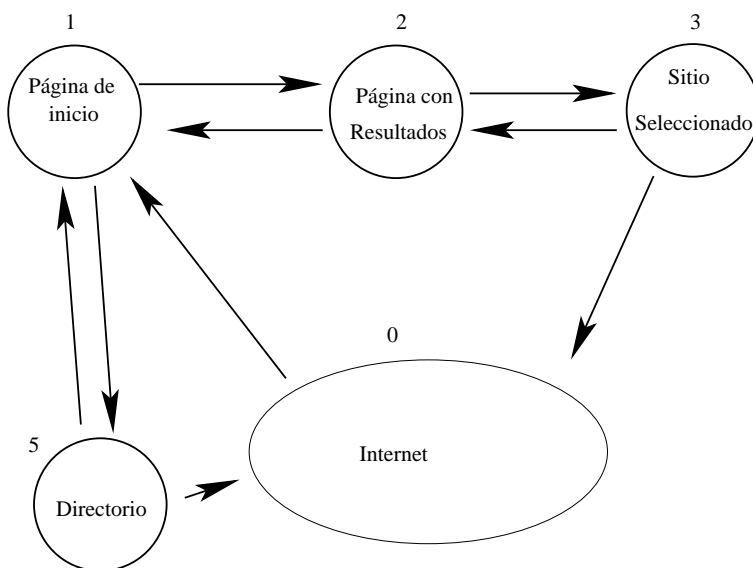


Figura 1: Mapa de navegación del sitio del buscador Todocl.

Una forma de representar la navegación es planteándola como un grafo, grafo que es conocido como el *mapa de navegación del sitio*. Para el caso del buscador Todocl, el mapa de navegación del sitio se muestra en la figura 1, indicando los estados posibles.

El análisis de las bitácoras genera el diagrama de la figura 2. Las transiciones entre estados (nodos) representan la porción de usuarios que al salir del nodo de origen fueron hacia el nodo de destino. Este número es una estimación de la probabilidad de que, dado que un usuario está en el origen del arco, vaya hacia el destino de éste. Dentro de cada estado está la probabilidad de que un usuario esté en esa página.

Hay varias observaciones que se pueden obtener de la figura 1.

- Los usuarios no usan las opciones avanzadas (menos del 1%).
- Es común el refinamiento de la consulta.
- Sólo el 9% de las personas usa el directorio de sitios.

Lo anterior indica que los usuarios, como método de búsqueda, utilizan la información del despliegue de los resultados. Es decir, un método más cercano a prueba y error que a una búsqueda más precisa, que sería el caso de la opción de búsqueda avanzada. Además al observar el grafo vemos que la probabilidad de que un usuario se encuentre en la página con resultados es mucho mayor que la probabilidad de que un usuario se encuentre en la URL seleccionada. Esto indica que los usuarios visitan relativamente pocas de las páginas del resultado como se mencionó anteriormente.



Palabra	cantidad de consultas	%
CHILE	10429	0.5%
FOTOS	9296	0.5%
GRATIS	8792	0.5%
SEXO	8061	0.4%
HISTORIA	7126	0.4%
MP3	4842	0.25%
VIDEOS	3675	0.2%
MUSICA	3376	0.2%
ARGENTINA	2941	0.15%
LEY	2803	0.15%
UNIVERSIDAD	2744	0.1%
VENTA	2658	0.1%
MEXICO	2610	0.1%
SOFTWARE	2549	0.1%
INTERNET	2492	0.1%

Tabla 1: Palabras más consultadas en Todocl.

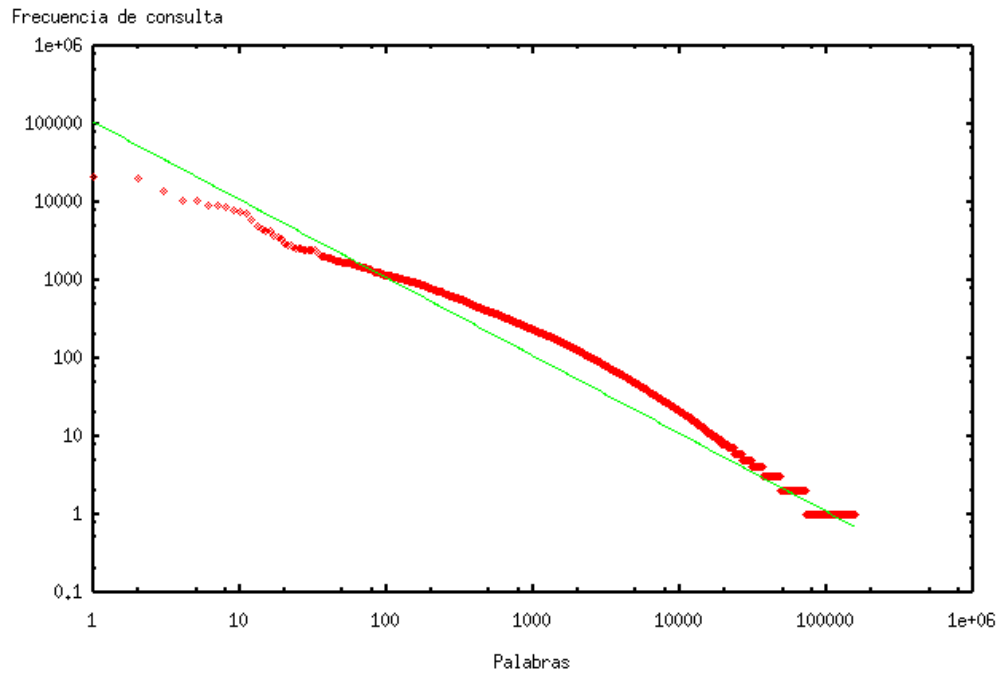


Figura 3: Frecuencia de las palabras consultadas en Todocl.

Opción	% de uso	frecuencia de uso
AND	99.9%	777.342
OR	0%	8
FRASE	0%	1
con acentos	0.1%	525
sin acentos	99.9%	777.021

Tabla 2: Uso de opciones de búsqueda en Todocl.

desviación estándar de la frecuencia diaria y la frecuencia total en el período, que se puede aproximar por el modelo  $x = 3y/2$ .

Algunas palabras tienen comportamientos particulares. Fuera del margen del gráfico en la figura 4 hay palabras que se salen de la escala por mucho, y se mantienen en la línea del modelo. Estas palabras son *de*, *en*, y *la*. Otra característica que es observable, es que aparecen algunas palabras de relativamente alta desviación pero baja frecuencia, por ejemplo, *boliviano* e *invierno*. *Invierno* es interesante ya que es una palabra claramente estacional, lo que justifica su baja frecuencia y alta desviación.

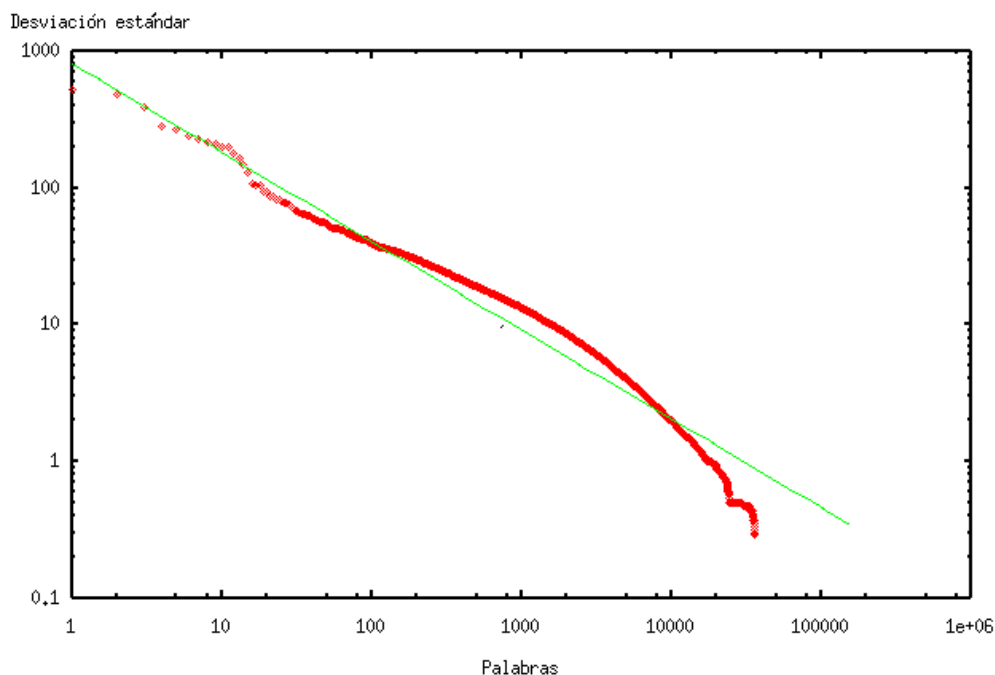


Figura 4: Desviación estándar de la frecuencia de las palabras consultadas en Todocl.

### 3.3 Opciones de Consulta

Al utilizar un buscador, es posible alterar los parámetros bajo los cuales se realizará la consulta. Los parámetros existentes en el modo de búsqueda simple, son:

**Operador** con valores AND, OR o FRASE. El valor AND busca documentos que tengan todas las palabras, OR documentos con alguna palabra, FRASE documentos que contengan la frase exacta.

**Acentos** considerar o no acentos en la consulta.

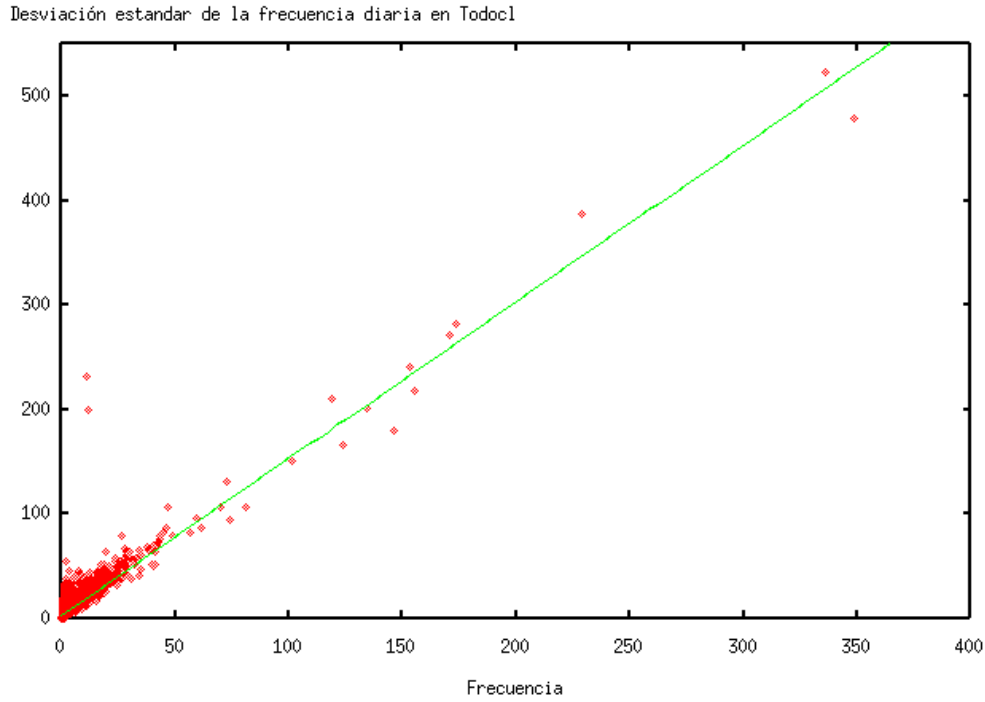


Figura 5: Frecuencia de consulta v/s desviación estándar en Todocl.

En la tabla 2 se pueden ver los niveles de utilización de cada opción en Todocl. Los valores más altos son los valores por omisión. Esto le da una tremenda importancia a las opciones por omisión, ya que su elección será determinante, en una gran cantidad de casos, para el buen resultado de las consultas.

### 3.4 Palabras Consultadas en el Contenido

Las palabras consultadas y las que aparecen en las páginas siguen distribuciones similares. Surge la pregunta sobre su relación. En el gráfico de la figura 6 se ve la relación entre documentos relevantes y cantidad de consultas de las palabras. Lo más común son palabras con pocos documentos relevantes y pocas consultas. Hay palabras con pocos documentos y muchas consultas, ejemplos de esto son *Hentai*, *México*, *DivX*, *carátulas* y *melodías*. Las palabras con muchos documentos relevantes y pocas consultas son, en general, palabras funcionales (llamadas *stopwords*). Por ejemplo, preposiciones, pronombres y artículos como *pero*, *otros*, *este*, etc. Las palabras con mucho contenido y muchas consultas son, en general, son también *stopwords* como *y*, *de*, *el* y *la*; pero aparece de forma interesante *Chile* como palabra muy consultada y que aparece en muchas páginas. Las palabras poco consultadas y con poco contenido no son interesantes, ya que son muchas. La relación entre las palabras consultadas y las del contenido no es clara, pero su correlación es baja.

## 4 Jerarquización de Páginas

En [6] se propone un algoritmo para mejorar los resultados de un buscador, basándose en las consultas hechas por los usuarios. El algoritmo propuesto es llamado MASEL (Matrix Analysis on Search Engine Log). Este algoritmo propone un método que relaciona consultas, usuarios y URLs seleccionadas por los usuarios. MASEL es propuesto para un buscador de multimedia, donde la comparación entre el documento y la consulta es de un nivel de precisión muy bajo.

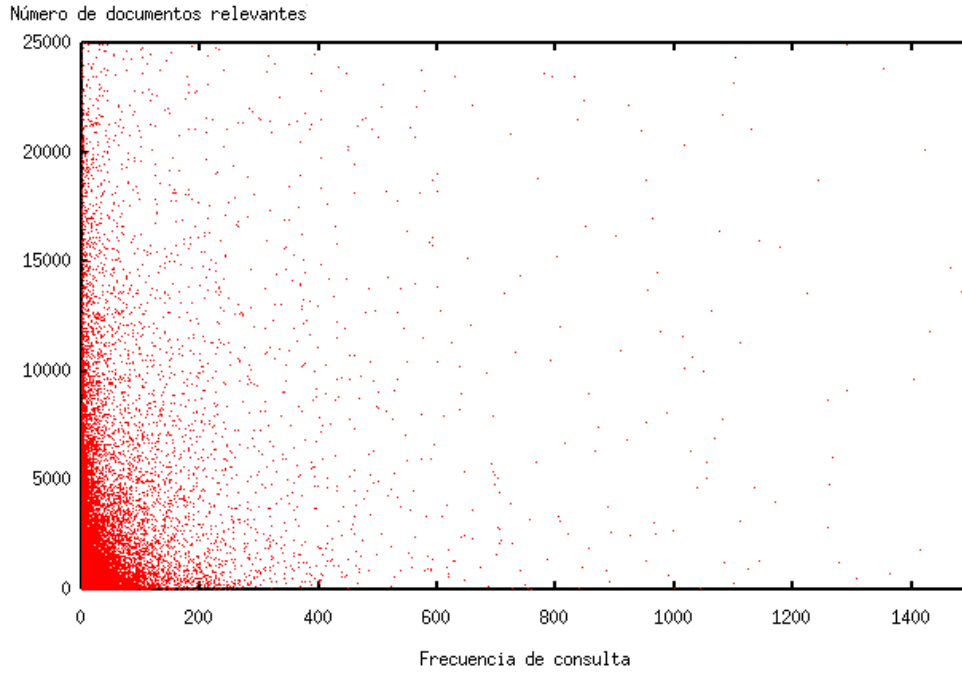


Figura 6: Cantidad de consultas v/s documentos relevantes para las palabras.

#### 4.1 El Algoritmo MASEL

La idea básica es obtener *recomendaciones* de usuarios que quedan reflejadas en los logs de acceso del buscador, ya que en ellos está la dirección IP del usuario, las consultas que hizo y las URLs que eligió. Es discutible cuanto identifica la dirección IP al usuario, pero se consideró que por un período corto de tiempo detrás de una IP está siempre el mismo usuario en la mayor parte de los casos.

El algoritmo considera simultáneamente consultas, usuarios y documentos, bajo una definición recurrente sobre estos tres elementos. Los buenos usuarios harán buenas consultas, las buenas consultas retornan buenos documentos y los buenos documentos son seleccionados por buenos usuarios.

Dada una consulta  $q^*$ , el algoritmo aplica cuatro etapas para obtener un conjunto de resultados:

- Primero se obtiene el conjunto de todos los usuarios que han consultado  $q^*$  recientemente, conjunto al que llamaremos  $U$ . Esta información se encuentra en los logs de acceso del buscador.
- Luego se obtiene el conjunto de todas las consultas realizadas recientemente por los usuarios en  $U$ , conjunto al que llamaremos  $Q$ . También en los logs de acceso del buscador está esta información.
- A continuación se obtiene, mediante técnicas tradicionales de recuperación de información (por ejemplo, usando el mismo buscador), el conjunto de documentos relevantes al conjunto de consultas  $Q$ . Este conjunto lo llamaremos  $R$ .
- Finalmente se calcula, mediante un proceso iterativo, los valores numéricos que permiten jerarquizar consultas, usuarios y documentos. Este proceso iterativo será explicado más adelante.

Al referirse a *recientemente* se está hablando del intervalo  $[t_{ahora} - T, t_{ahora}]$  donde  $T$  es un parámetro del algoritmo. Sean  $m = |U|$ ,  $s = |R|$  y  $n = |Q|$ , además sean  $u_i$  el valor que representa la calidad del usuario  $i$ ,  $q_j$  la calidad de la consulta  $j$  y  $r_k$  la calidad del documento  $k$  (luego se explicará cómo se obtienen estos valores). Utilizaremos valores normalizados, es decir

$$\sum_U u_i = 1, \quad \sum_Q q_j = 1, \quad y \quad \sum_R r_k = 1$$

lo que no afecta los resultados, ya que nos interesa el orden de ellos, y no los valores mismos de calidad. Esta normalización debe realizarse luego de cada iteración.

Definimos los pesos para los usuarios según sus buenas consultas:

$$u_i = \sum_{j=1}^n a_{ij} q_j, \quad \text{donde } a_{ij} = \begin{cases} num(u_i, q_j) & \text{si } q_j = q^* \\ \alpha num(u_i, q_j) & \text{si } q_j \neq q^* \end{cases}$$

donde  $num(i, j)$  representa cuántas veces el usuario  $u_i$  realizó la consulta  $q_j$  recientemente y  $0 < \alpha < 1$  es un peso para quitar relevancia a las consultas distintas a  $q^*$ .

Definimos las buenas consultas según recuperen buenos documentos:

$$q_j = \sum_{k=1}^s b_{jk} r_k, \quad \text{donde } b_{jk} = \begin{cases} sim(q_j, r_k) & \text{si } q_j = q^* \\ \beta sim(q_j, r_k) & \text{si } q_j \neq q^* \end{cases}$$

donde  $sim(q_j, r_k)$  es la similitud entre el documento  $r_k$  y la consulta  $q_j$  y  $0 < \beta < 1$  es un peso para quitar relevancia a las consultas distintas a  $q^*$ . Así definimos los buenos documentos según la preferencia de los buenos usuarios:

$$r_k = \sum_{i=1}^m c_{ki} u_i, \quad \text{donde } c_{ki} = hit(r_k, u_i, \{q^*\}) + \gamma hit(r_k, u_i, Q - \{q^*\})$$

donde  $hit(r_k, u_i, S)$  es la cantidad de veces que el usuario  $u_i$  eligió el documento  $r_k$  al hacer una consulta en el conjunto  $S$  y  $0 < \gamma < 1$  es un peso para quitar relevancia a las consultas distintas a  $q^*$ . En forma matricial tenemos

$$u = Aq, \quad q = Br, \quad r = Cu$$

lo que implica que

$$u = Aq = A(Br) = A(BCu) = (ABC)u$$

$$q = Br = B(Cu) = B(CAq) = (BCA)q$$

$$r = Cu = C(Aq) = C(ABr) = (CAB)r$$

Esto define un método iterativo para calcular  $u$ ,  $q$  y  $r$ . Es decir, podemos obtener los mejores usuarios, consultas y documentos.

El parámetro  $T$ , que define la ventana de tiempo en la cual miramos las bitácoras, es fundamental en la efectividad del algoritmo. Si  $T$  es muy grande, el conjunto de consultas extendido comienza a abarcar demasiados elementos no relevantes a la consulta inicial. Por otro lado, si  $T$  es muy chico, estamos desperdiando información. Lo complejo de la determinación del parámetro  $T$  es que su valor debe depender de la frecuencia de la consulta (que ya vimos sigue una distribución de Zipf). Además un parámetro  $T$  mayor hace las matrices más grandes y el algoritmo más lento.

El método operó bien en Todocl, dando resultados relevantes a las consultas. En general el resultado tiende a un par de documentos con calificación mucho mayor que el resto, lo que hace de MASEL un buen algoritmo para utilizarlo como una ayuda paralela a los resultados del buscador entregando al usuario sitios populares relacionados a su consulta. Por ejemplo se consultó por *autos* con  $T = 24$  horas y  $T = 48$  horas. Con  $T = 24$  horas las URLs más recomendadas fueron:

<http://www.chileautos.cl/link.asp>  
<http://www.chileautos.cl/personal.htm>  
<http://www.autoscampos.cl/frmencabau.htm>  
<http://rehue.csociales.uchile.cl/publicaciones/moebio/07/bar02.htm>

Las primeras tres páginas son relevantes, pero la cuarta no. Con  $T = 48$  horas las URLs más recomendadas fueron:

<http://www.mapchile.cl/>  
<http://fid.conicyt.cl/acreditacion/normas.htm>  
[http://www.clancomunicaciones.cl/muni\\_vdm/consejo.htm](http://www.clancomunicaciones.cl/muni_vdm/consejo.htm)



donde la segunda y tercera URL no son relevantes, pero la primera es relevante sin contener la palabra *autos*, al ser un sitio de mapas que se puede considerar útil para un automovilista. Esto muestra la capacidad del algoritmo de relacionar semánticamente recursos de la Web.

Podemos ver en este ejemplo la muy alta dependencia del método al parámetro  $T$ .

## 4.2 Análisis de Precisión

Se aplicó el algoritmo anterior a palabras de distinto nivel de frecuencia de consulta para distintos valores del parámetro  $T$ . En la figura 7 se ve la precisión del algoritmo para las consultas *software*, *bancos* y *empresas*. Las tres palabras son frecuentemente consultadas. *Software* es bastante más consultada que las otras dos, y vemos que el algoritmo comienza a obtener buenos resultados para valores de  $T$  menores. *Banco*, para  $T = 72$ , obtiene una excelente precisión, la que decrece al aumentar  $T$ . Es claro que para un buen funcionamiento del algoritmo,  $T$  debe depender de la frecuencia de consulta de la palabra, y no ser constante como propone el trabajo original. Cabe mencionar que donde la precisión es 0 es porque el algoritmo no retornó resultados.

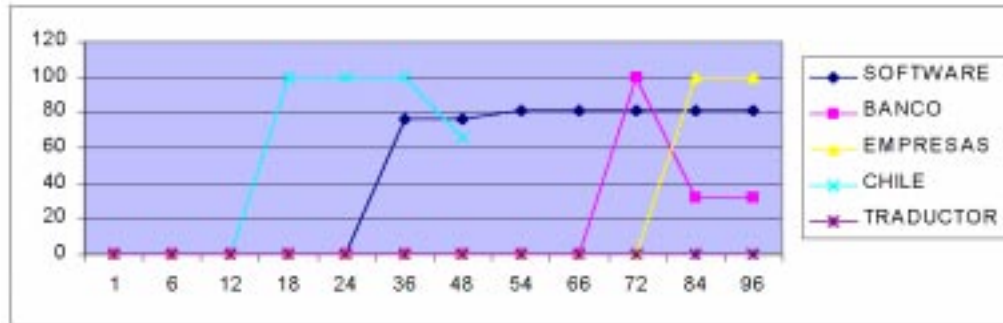


Figura 7: Parámetro  $T$  v/s precisión del algoritmo.

Una observación interesante es que en algunas de las consultas realizadas entre los resultados surgía un tema secundario con gran cantidad de documentos relevantes. Por ejemplo, para la consulta *banco*, muchos de los documentos no relevantes se referían a venta de casas. Otro ejemplo es que en la consulta *software* surgió como tópico secundario el reciclaje de vidrios, siendo casi todos los documentos no relevantes relacionados a este tema. Este fenómeno se llama desplazamiento temático (*topic drifting*) en recuperación de información en la Web cuando se usa información de enlaces entre páginas.

Para consultas aún más populares los tamaños de las respuestas fueron lo suficientemente grandes como para hacer imposible el análisis de los datos en forma manual, por lo que no se pudo determinar la relevancia de los resultados. Este fue el caso de las consultas *audio* y *moda*. Si la frecuencia de la consulta no es alta, no es posible usar esta técnica, por ejemplo para la consulta *ropa* no surgieron documentos relevantes, y para *traductor* se recuperaron siempre los mismos dos documentos, ambos no relevantes.

Para la consulta *chile*, en pocas de las páginas Chile era el tema principal, siendo en casi todas un tema relacionado. Por otro lado, el tema principal en todas las otras páginas fue relativo a celebridades o deportes en Chile, lo que es interesante ya que al hacer una búsqueda por Chile no aparece ningún resultado relativo a estos temas en los primeros 30 resultados. Para valores de  $T$  mayores a 48 la cantidad de documentos recuperados fue enorme.

## 5 Conclusiones

Hemos presentado un análisis de consultas de un buscador chileno, el cuál da información de los internautas en Chile. Resumiendo, podemos decir que las consultas no son sofisticadas y no aprovechan todo el potencial del buscador. Por otra parte, es necesario ahondar en el análisis y descubrir si hay patrones en las consultas. Por ejemplo, palabras permanentes, palabras con una cierta periodicidad y palabras transitorias. Nuestro

análisis de varianza no permitió distinguir entre las palabras permanentes y transitorias, al parecer porque la frecuencia de estas últimas es muy baja.

Respecto al uso de las consultas para *ranking*, es necesario definir la dependencia de  $T$  en la frecuencia de la palabra, de manera de obtener buenos resultados a partir de todas las consultas, ya que el algoritmo es poco eficaz para valores muy pequeños de  $T$  y poco eficiente (además de poco eficaz) para valores muy grandes de  $T$ . Un análisis más detallado podría dar un modelo aproximado para la dependencia de  $T$  con respecto a la frecuencia.

Por el tipo de resultados del algoritmo es recomendable su utilización como ayuda complementaria a una búsqueda con métodos más estables de recuperación de documentos, debido a que para muchas consultas poco populares recupera pocos o ningún documento. En términos de interfaz al usuario, además de los resultados tradicionales de la búsqueda, se recomienda agregar un cuadro lateral que mencione recursos populares entre los usuarios con los mejores resultados de MASEL frente a la consulta realizada.

Nuestros resultados indican que el problema de generalización del tema o desplazamiento del mismo (*topic drifting*) también aparece al usar información de navegación. En nuestro es posible que su aparición pueda ser controlada mediante la reducción de los valores fijados para los parámetros  $\alpha$ ,  $\beta$  y  $\gamma$ .

Los resultados también podrían ser mejorados detectando distintas sesiones para el mismo número IP, las que pueden indicar distintos usuarios. Sin embargo, detectar sesiones en forma precisa es uno de los problemas más difíciles del análisis de bitácoras.

## Referencias

- [1] Doug Beeferman and Adam Berger. Agglomerative clustering of a search engine query log, Proceedings on the 2000 Conference on Knowledge Discovery and Data Mining (Boston,MA), pages 407-416, Aug. 2000
- [2] Craig Silverstein, Monika Henzinger, Hannes Marais, and Michael Moricz. Analysis of a Very Large AltaVista Query Log. *SIGIR Forum* 33(3), 1999.
- [3] Amanda Spink, Bernard J. Jansen, Dietmar Wolfram, Tefko Saracevic. From E-Sex to E-Commerce: Web Search Changes. *IEEE Computer* 35(3): 107-109 (2002).
- [4] TodoCL: Home Page, [www.todocl.cl](http://www.todocl.cl), 2000.
- [5] Dietmar Wolfram. A Query-Level Examination of End User Searching Behaviour on the Excite Search Engine. Proceedings of the 28th Annual Conference Canadian Association for Information Science, 2000.
- [6] Dell Zhang and Yisheng Dong. A Novel Web Usage Mining Approach For Search Engine. Por aparecer en *Computer Networks*, 2002.