

ANÁLISIS DE CONSULTAS A UN BUSCADOR DE LA WEB CHILENA

Ricardo Baeza-Yates¹

Carolina Galleguillos²

RESUMEN

Los buscadores de páginas Web son herramientas indispensables de orientación para sus usuarios, lo cual refleja la gran cantidad de personas que los utilizan diariamente. Las bitácoras (*logs*) que registran el uso de un buscador son una valiosa fuente de información: nos hablan de las expectativas, preferencias y necesidades de las personas, es decir, que esperan ellas encontrar en la Web. El resultado del análisis de logs de uso tiene un importante valor comercial, en particular para el comercio electrónico y la publicidad en Internet.

El objetivo de este estudio es encontrar relaciones entre las consultas de los usuarios de la Web Chilena (.cl) y las diversas áreas de clasificación de sus diferentes sitios. También se espera encontrar relaciones entre los tipos de búsqueda realizados y los sitios visitados, para así conocer la necesidad que los lleva a formular sus consultas. La información obtenida está basada en los logs del Buscador Chileno TodoCL entre Febrero y Marzo del 2004.

Palabras Claves: minería de uso, Web chilena, comportamiento de búsqueda de usuarios en la Web, clasificación de consultas.

ABSTRACT

Web page search engines are essential tools for users' orientation, which is reflected by the big amount of people who are using them every day. The logs which reflect a Web search engine uses are a valuable source of information: they give us information about the users' expectations, preferences and needs, which is everything that they expect to find on the Web. The results from the logs usage analysis have an important commercial value, in particular, for e-commerce and Internet publicity.

The purpose of this study is to find relations between Chilean Web users' queries (.cl) and the different classification areas for its web sites. Also we expect to find relations between the search types made and the web sites visited, trying to get the need behind the query made. The data used for this study was gotten from the Chilean search engine TodoCL, between February and March 2004.

Keywords: Web usage mining, Chilean Web, Web users search behavior, query classification.

¹ Universidad de Chile, Departamento de Ciencias de la Computación, Blanco Encalada 2120, Santiago 6511224, Chile, rbaeza@dcc.uchile.cl

² Universidad de Chile, Departamento de Ciencias de la Computación, Blanco Encalada 2120, Santiago 6511224, Chile, cgallegu@dcc.uchile.cl

INTRODUCCIÓN

La Web es un gran repositorio de información que va cambiando constantemente, debido a su crecimiento exponencial y a su interacción diaria con millones de personas. Detrás de ella existe un complejo mundo de relaciones que a simple vista es difícil de distinguir. Esta información refleja los gustos y preferencias de los usuarios, sus intereses, sus personalidades. En este estudio la Web Chilena es un pequeño subconjunto de la Web, pero no es una muestra de la Web global debido a que los sitios de un país son muchos más homogéneos, comparten un solo idioma, una sola cultura y el contexto es común en la mayoría de los casos [2]. De esta forma nuestra fuente de información es mas bien una colección de páginas bien formada.

La minería de la Web (*Web mining*) es la herramienta que nos permite encontrar este tipo de relaciones. Podemos definir Web mining como la extracción de información relevante desde la Web Global (World Wide Web) seguida de su generalización y análisis [3]. De esta forma podemos encontrar diferentes tipos de Web mining: Web mining de Contenido, que incluye el minado de recursos disponibles en los sitios Web y en las bases de datos online (texto y multimedia) [4]. Web mining de estructura, que implica el análisis de la estructura de enlaces de la Web [5]. Web mining Temporal o Dinámico, que se enfoca en el crecimiento y cambios de la Web [6]. Web mining de Uso, consiste en el descubrimiento y análisis de patrones de acceso de usuarios desde uno o más servidores Web o servicios online (minado de logs) [4]. Este último tipo de Web mining es la herramienta principal usada en este estudio, para la obtención de datos y descubrimiento de relaciones, a través del minado de logs del buscador.

El propósito de este estudio es encontrar relaciones entre los distintos tipos de áreas en las cuales se clasifican los sitios pertenecientes a la Web Chilena (.cl), con respecto al comportamiento de los usuarios en la búsqueda de sitios Web, utilizando Web mining de uso.

Este estudio se organiza de la siguiente manera. Se comenzará presentando las características de los datos usados para el estudio y el escenario global que presenta la Web Chilena. A continuación de esto, se presentan las distintas categorías en las cuales se han clasificado los sitios Web y su relación con los tipos de búsqueda en la Web. Luego, se presentarán las estadísticas encontradas en el análisis de los logs, el comportamiento de cada una de las áreas en los distintos meses y entre ellas, y sus consultas. Finalmente se presenta la conclusión acerca de los

resultados obtenidos, los que además de tener valor técnico, tienen sin duda un valor comercial.

ALGUNOS DATOS SOBRE LA WEB CHILENA

En esta sección presentamos algunas estadísticas globales de la Web Chilena. Esto nos permite tener una idea general del escenario en el cual los usuarios buscaron durante el período en el que se recolectó la información. También se presenta la clasificación en áreas, para los distintos sitios que componen la Web Chilena.

Estadísticas Generales

Para los experimentos se consideraron los logs del buscador TodoCL, el cual está enfocado principalmente en páginas de la Web Chilena. TodoCL es un motor de búsqueda de indexación automática, que además incorpora un directorio (usa ODP: www.dmoz.org). Su colección se compone actualmente de alrededor de 3.100.000 páginas Web chilenas, además de archivos en distintos formatos como Pdf, PostScript y Word. La Web Chilena está compuesta por todos los dominios .cl, y algunos .net y .com hospedados en proveedores de Internet chilenos [7].

La muestra de logs corresponde al período entre el 01 Febrero 2004 - 00:00 y el 31 Marzo 2004 - 23:59. Se analizaron un poco más de 5 millones (5.142.351) de transacciones (alrededor de 1GB), en los cuales se encontraron alrededor de 758 mil URLs diferentes. El total de URLs encontradas considera las URLs internas del sitio de búsqueda más las URLs de salida y URLs de entrada. De este total de URLs sólo 23.307 corresponden a URLs de sitios Web chilenos a los cuales les estaba asociado a una consulta en el buscador (URLs de salida). De estos últimos, 22.795 eran sitios con dominio .cl, registrados en NIC Chile durante el 2003[8].

	Febrero	Marzo
Visitantes únicos	153.773	172.917
Número de visitas	272.010	314.964
Páginas visitadas	1.121.763	1.259.525
Hits	1.408.670	1.623.350
URLs únicas	377.989	418.771
URLs usadas (.cl)	11.314	11.481

Cuadro 1: Estadísticas Básicas TodoCL.

Todas estas estadísticas, correspondientes a los meses de Febrero y Marzo 2004, excluyen los logs generados por los crawlers de los distintos buscadores existentes. Podemos ver en el Cuadro 1, que la cantidad de visitantes en Marzo es mayor que en el caso de Febrero, debido principalmente a que Febrero es un mes donde generalmente la gente sale de vacaciones a otros lugares o dedica su tiempo a otras actividades, en cambio Marzo es un mes activo, en el cual la gente vuelve a su actividades que realizan normalmente durante el año.

De la misma forma se puede ver en las páginas visitadas, donde Marzo es mayor que Febrero, y asimismo el resto de los indicadores. Sin embargo, si bien la cantidad de URLs únicas de Marzo es un poco mayor que aquellas usadas en el mes de Febrero, la proporción de URLs usadas son casi las mismas para los dos meses. Esto significa que los usuarios que llegaron al buscador en Febrero generaron casi la misma cantidad de consultas al buscador que en Marzo, y utilizaron menos el directorio. Esto puede deberse a que en el mes de Marzo es un mes donde el consumo de bienes aumenta, dado el comienzo de clases para los estudiantes, con lo que la necesidad de los usuarios es principalmente bienes y servicios, que son más intuitivos de buscar en el directorio. En cambio Febrero es un mes de vacaciones, lo que implica que la gente se comporta de una manera diferente frente a la Web, sin una tendencia marcada.

Con respecto al tipo de usuarios que visitaron al buscador dentro del período, se puede observar en las Figuras 1 y 2 que si bien la cantidad de días en los dos meses es igual, el porcentaje de usuarios por país cambió. En ambos meses los 10 primeros países son los mismos y en general conservan su proporción de visitas. También se observa que los países son en su mayoría de habla hispana, y aquellos que no lo son su idioma tienen palabras que son comunes y usadas en el español, como es el caso del portugués (Brasil). Aquellos países que tienen como idioma el español son países que tienen gran cantidad de usuarios de la Web Global y de la Internet en general. Podemos también observar que la mayoría de los usuarios tiene IP desconocida o sin resolver, y se ven usuarios con .net (que generalmente corresponden a empresas) y .cl, correspondiente a usuarios chilenos.

De esta manera se puede apreciar que el escenario de los usuarios es prácticamente el mismo, son siempre visitantes de los mismos países, por lo que el contexto es el mismo. Lo que cambia sin duda es el propósito con que los usuarios realizan su búsqueda, y la forma en que interactúan con la Web.



Figura 1: Los 10 dominios de visitantes con más hits en Febrero.

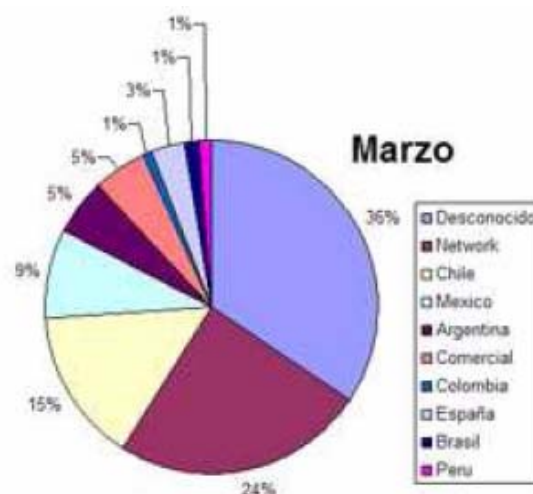


Figura 2: Los 10 dominios de visitantes con más hits en Marzo.

Clasificación de los sitios Chilenos

Detrás de cada consulta hay una necesidad. El comportamiento de los usuarios en la Web permite distinguir 3 tipos de clases de consultas [9]: de navegación, de información y de tipo transaccional. Las primeras corresponden a la necesidad de encontrar un sitio en particular. Las de información, que como lo dice su nombre intentan adquirir información asumiendo que existe en uno o más sitios Web. En la de transacción, el propósito es realizar alguna actividad dada por los servicios ofrecidos de un sitio Web.

Actualmente los sitios chilenos que se registran bajo el dominio .cl en NIC Chile [8] se clasifican en las siguientes categorías:

1. **Comercio:** Sitios correspondiente a empresas. Generalmente se ofrecen productos, servicios o información de la empresa.
2. **Educación:** Sitios correspondientes a instituciones relacionadas con la educación, instituciones de educación superior y media, y proyectos educacionales.
3. **Gobierno:** Sitios relacionados y dependientes del gobierno, partidos políticos.
4. **Militar:** Sitios relacionados con las fuerzas armadas.
5. **Otros:** Sitios independientes, páginas personales, hobbies, fans club, organizaciones sociales, agrupaciones.
6. **Particular:** Sitios independientes que corresponden a empresas pequeñas, asociaciones, organizaciones sociales y grupos con fines comunes.

Es posible distinguir ciertas relaciones entre las clases de búsqueda y las categorías de clasificación. Debido a la naturaleza de la clasificación de los sitios, aquellos sitios correspondientes al área comercio pueden satisfacer principalmente búsquedas de navegación y transacción en su mayoría. Para educación, son en general búsquedas de información. En el caso de “gobierno” es información y navegación. Para “militar”, mayormente información. El área de clasificación “otros” es de navegación e información. Finalmente el área “particular” tiene transacciones e información.

Pero cabe destacar que las búsquedas de información de área de comercio son distintas a aquellas búsquedas de información que un usuario puede hacer para sitios clasificados en el área de gobierno. Dentro de las categorías de búsqueda de Información y Transacción, existen diferentes subcategorías [10] dentro de ellas, las cuales permiten tener una idea de una necesidad específica de búsqueda por parte del usuario.

Las subcategorías, presentadas por Rose y Levison [10] para las búsquedas de información son: Directas, cuando se quiere encontrar algo específico de un tema (con respuestas abiertas o cerradas), Indirectas (cuya respuesta puede ser cualquier cosa relacionado con el tema que pregunto), de Sugerencia, de Localización (de algún producto o servicio) y de Listado (cuando se busca tener una lista de sitios Web). Para las búsquedas de tipo transaccional o de recursos (no información), pueden ser divididas en: Descarga (de algún recurso), Entretenimiento, Interacción (con algún recurso o servicio) y Obtención (que puede ser obtenida en la pantalla o para ser impresa). En el caso de la búsqueda

de tipo Navegación, no existen subcategorías, ya que de por si el objetivo es específico.

LA NECESIDAD DETRÁS DE LA CONSULTA

En esta sección se mostrará toda la información obtenida acerca de las consultas realizadas por los usuarios en el buscador dentro del período descrito anteriormente. Se presentará n las consultas agrupadas por las distintas áreas de clasificación y se verá como cambian entre ellas y con respecto a los distintos meses.

Análisis de Logs

Como se describió anteriormente, se ocuparon alrededor de 22795 URLs tónicas (URLs de salida), las cuales estaban asociadas a una consulta, realizada en el buscador TodoCL. Además de estas URLs, se contaba con otra lista de URLs, donde cada una estaba asociada a un área o categoría (comercio, educación, gobierno, militar, otros, particular). De esta forma, se relacionaron las consultas con las distintas áreas.

Se analizaron las consultas para cada mes en particular, y después se buscaron todas las consultas en común entre los dos meses para comparar el cambio de las frecuencias de cada una de las áreas.

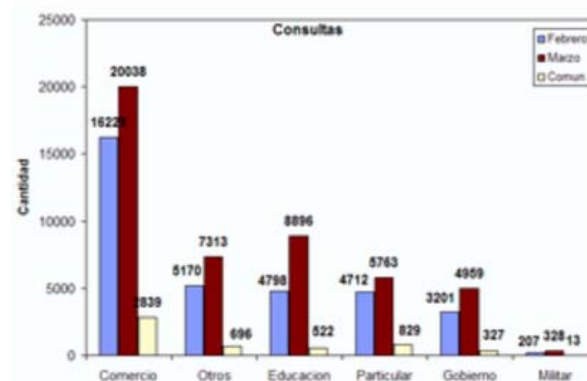


Figura 3: Cantidad consultas distintas para Febrero y Marzo.

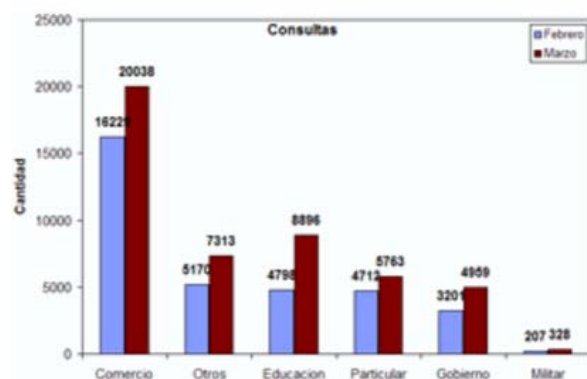


Figura 4: Cantidad consultas iguales para Febrero y

Marzo.

En las figuras 3 y 4 podemos ver que como era de esperarse, el mes de Marzo tiene una mayor cantidad de consultas con respecto a Febrero. La figura 3 nos muestra que la cantidad de consultas en común entre los meses es bastante menor que las cantidades de los meses por si solos. Esto indica que son pocas las consultas que se repiten por mes, y que los usuarios están generando nuevas consultas todo el tiempo.

Vemos que en general, las áreas no varían mucho su comportamiento con respecto al tiempo. Es comercio el área que tiene más cantidad de consultas, y mayor frecuencia de las consultas. Una de las razones por la cual esto ocurre, es el hecho de que existen mayor cantidad de sitios Web que pertenecen a esta área. El área con menores consultas es el área militar, probablemente debido a que existen pocos sitios dentro del área y poca gente interesada en ellos. El resto de las áreas cambian un poco, como por ejemplo educación, que obtiene menos consultas para Febrero, debido claramente a que es un mes de vacaciones para los estudiantes. En su contraparte para el área otros, la cantidad de consultas aumenta, ya que relaciona actividades de esparcimiento y diversión.

En las figuras 5 y 6 podemos ver que comercio y militar son las áreas opuestas con respecto a las frecuencias de las consultas en ambos meses. Para el resto de las áreas, con respecto al mes de Marzo el área de educación estaba en el segundo lugar, pero en el mes de Febrero estaba el área particular, para consultas diferentes (figura 5). En la figura 6 se observa que las áreas siguen las mismas tendencias y las cantidades entre los distintos meses son muy parecidas.

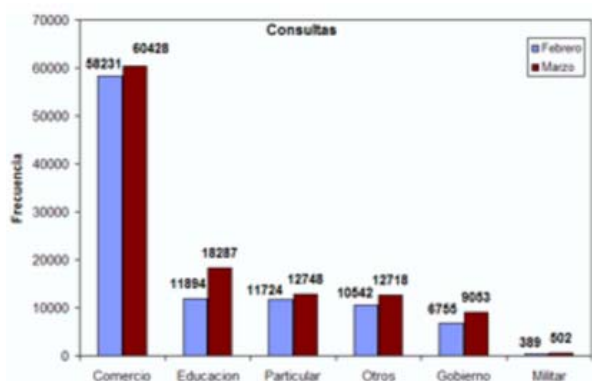


Figura 5: Frecuencia consultas distintas para Febrero y Marzo.

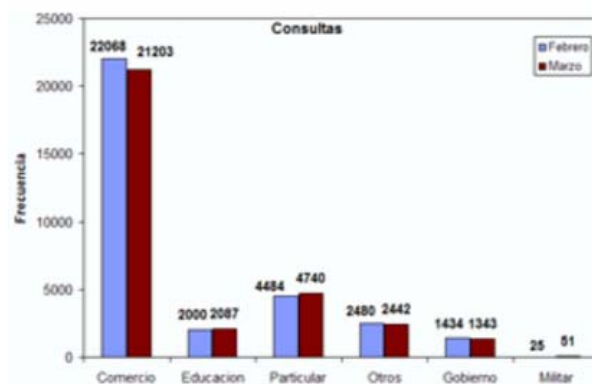


Figura 6: Frecuencia consultas iguales para Febrero y Marzo

Distribuciones

A través de los modelos de distribuciones es posible apreciar de mejor forma el comportamiento de las áreas para cada una de las consultas. Las siguientes figuras muestran los modelos de las distribuciones para cada una de las áreas, para cada mes.

En las figuras 7 y 8 podemos observar que las frecuencias de todas las áreas siguen una distribución de Zipf [11]. Para cada curva su primera parte corresponde a consultas que son más frecuentes, en cambio la última parte corresponde a consultas únicas, con prácticamente una sola aparición. La curva correspondiente al área comercio es muy parecida a la curva que representa a todas las áreas en ambos meses.

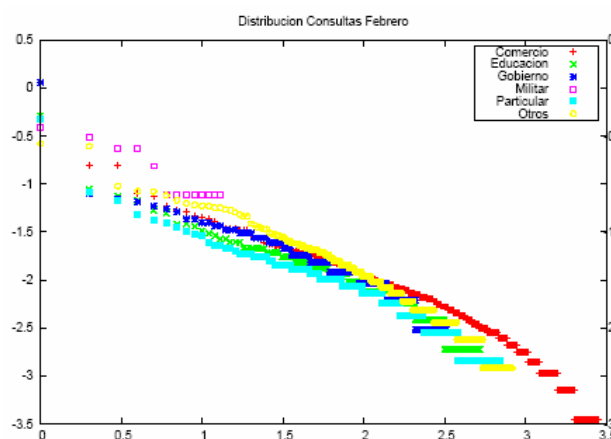


Figura 7: Frecuencia normalizada de todas las áreas para Febrero.

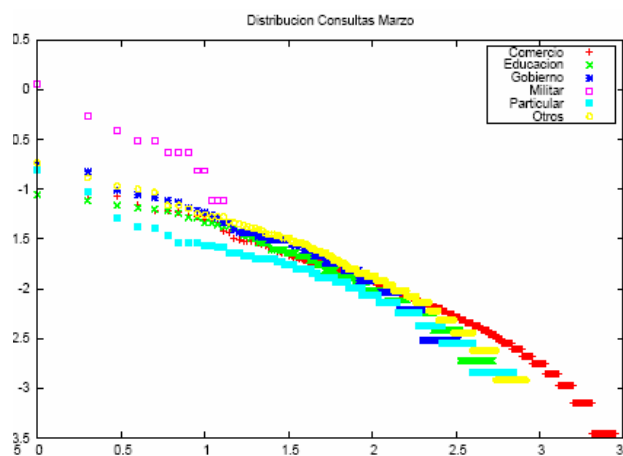


Figura 8: Frecuencia normalizada de todas las áreas para Marzo.

Área	Pendiente
Todas	1.10815
Comercio	1.08484
Educación	0.85770
Gobierno	0.81455
Militar	0.73424
Otros	0.82577
Particular	1.05796

Cuadro 2: Pendientes Distribuciones Diciembre.

Área	Pendiente
Todas	0.94557
Comercio	0.89943
Educación	0.70768
Gobierno	0.68421
Militar	0.39187
Otros	0.64967
Particular	0.86618

Cuadro 3: Pendientes Distribuciones Enero

Podemos observar las pendientes de las distribuciones en los Cuadros 2 y 3. Si comparamos en función del tiempo, vemos que las distribuciones cambian en ambos meses para cada área. En el caso de las distribuciones para comercio y particular, podemos observar que para cada mes por separado son bastante parecidas entre sí. En el caso de las demás áreas vemos que educación, otros y gobierno son similares pero sólo en el mes de Febrero.

La figuras 9 y 10, nos muestran el comportamiento para las consultas comunes en ambos meses. Podemos observar que para el área de comercio no parece haber alguna relación estable entre los dos meses para sus consultas. Esto quiere decir que algunas consultas se repiten en ambos meses con frecuencias diferentes. Con respecto al área gobierno, podemos ver que hay muy poca estabilidad, y no es posible ver a simple vista si hay frecuencias proporcionales entre meses para algunas consultas. De todas las áreas, militar fue el área con menor estabilidad, probablemente debido a la poca cantidad de consultas comunes (sólo 13). Sin embargo, casi todas las áreas (exceptuando el área referente a las consultas de sitios militares) se parecen entre sí, en su comportamiento con respecto a las frecuencias entre los dos meses. Todas tienen una gran cantidad de consultas con muy poca frecuencia y pocas con grandes

cantidades, pero no muestran una clara relación entre ellas.

Consultas

Con respecto a las consultas encontradas en ambos meses, podemos decir que de un total de 51.132 consultas efectuadas en el buscador dentro de Febrero y Marzo, 12.359 consultas estaban compuestas por sólo una palabra. De esta manera, el promedio de palabras por consulta es alrededor de 2.6 para todo el período de experimentación.

Otra característica importante de destacar, es que los sitios con los cuales están relacionadas estas consultas están influenciados de una manera u otra por el algoritmo de ranking del buscador. Esto hace que muchos sitios no sean visitados por el hecho de aparecen después de los 10 primeros resultados. Las consultas más populares dentro de los 2 meses fueron: casas prefabricadas, porno y autos usados. Todas ellas aparecieron en las áreas de comercio, particular y otros.

N	Consulta	Porcentaje
1	sexo	57.81%
2	autos+usados	84.18%
3	prefabricadas	81%
4	fotos	56.33%
5	postales	70.89%
6	acompañantes	86.09%
7	loteria	98.92%
8	cortinas	87.50%
9	kino	100%
10	lenceria	61.83%
11	cabañas	79.21%
12	motos	86.81%
13	sap	96.30%
14	porno	33.77%
15	parcelas	77.78%
16	juguets	86.36%
17	tierra	90.24%
18	bicicletas	82.22%
19	arriendos	80.72%
20	escorts	60.19%

Cuadro 5: 20 consultas más frecuentes área comercio Diciembre 2003.

N	Consulta	Porcentaje
1	sexo	35.06%
2	porno	35.50%
3	masajes	48.70%
4	chat	43.65%

5	gay	44.44%
6	culitos	84.48%
7	colegialas	69.01%
8	escort	48.04%
9	lenceria	32.06%
10	fotos	16.33%
11	saunas	75%
12	escorts	33.98%
13	sex	38.96%
14	amateur	72.97%
15	chicas	28.89%
16	postales	14.56%
17	moteles	25.61%
18	culos	30.30%
19	casas+prefabricadas	11.17%
20	chilenas+desnudas	70.37%

Cuadro 6: 20 consultas más frecuentes área particular Diciembre 2003.

Los Cuadros 5 y 6 nos muestran las consultas más frecuentes dentro de las áreas comercio y gobierno. El porcentaje que corresponde a cada una es el porcentaje de consultas relacionadas con URLs pertenecientes a esas áreas, con respecto al total encontrado en el mes de Marzo. Como podemos observar, las áreas comercio y particular tienen varias consultas en común, lo que puede deberse a que ambas están compuestas en su mayoría por empresas. Esto hace más difícil poder distinguir consultas específicas para cada una de estas áreas. Vemos en las tablas anteriores que la necesidad de los usuarios es principalmente productos, como autos usados, muebles, llantas, y servicios, como pilates y abogados, en el caso de comercio. Sin embargo para particular, la necesidad es mucho menos específica, debido quizás a que no es posible situar en un contexto las consultas.

En el caso de las áreas gobierno, educación y militar, se pudieron observar consultas que son representativas de ellas mismas. Por ejemplo para educación las consultas más populares fueron: universidad, universidades, diplomados. Para gobierno se encontró: código del trabajo, leyes, registro+civil. El área militar arrojó como las consultas más populares: carabineros, escuela militar, armada de chile. Para el área de clasificación otros, se encontraron los más diversos tipos de consultas. Entre las más populares están: sexo, porno, propiedades, camping, ong. Debido principalmente a la naturaleza de los sitios en esta área, será muy difícil encontrar consultas representativas, ya que los sitios no tienen un propósito definido y es posible encontrar diferente tipo de información y servicios en ellos.

Es importante considerar que los tipos de consultas formuladas en el buscador por los usuarios están fuertemente ligadas a los acontecimientos que ocurran en la sociedad. El comportamiento de los usuarios está

definido por lo que pasa a su alrededor. Así vemos que en Marzo dentro de las consultas más frecuentes se encontraron “casas+prefabricadas”, “autos+usados”, “arriendos” y “propiedades”. Para el mes de Febrero, mes de vacaciones de verano en Chile, encontramos “arriendos”, “chile”, “mapas de chile” y “parcelas”.

Para el área comercio, podemos distinguir búsquedas de información de tipo: indirecta, de localización y de listado, de navegación y de transacción de tipos: de descarga y de interacción. Para el área de educación se encuentran búsquedas de navegación, y de información de tipo indirecta y listado. Para gobierno hay búsquedas de navegación y de información de tipo directa (abierta) e indirecta. En el caso del área militar se distinguen búsquedas de navegación y de información directa (abierta). En el área otros las búsquedas de navegación, información y transacción están presentes. En el caso de información encontramos de tipo indirecta y listado, y para transacción, descarga y obtención. Finalmente para particular, tenemos búsquedas de navegación, de información de tipo indirecta, de listado y localización, y de transacción de tipo obtención.

CONCLUSIÓN

En este estudio nuestra intención fue mostrar la existencia de relaciones entre las áreas de clasificación de los sitios Web Chilenos y las consultas que los usuarios generan para llegar a esos sitios.

Con respecto a los usuarios, podemos decir que son uniformes con respecto al tiempo, ya que la mayoría de los usuarios que visitan el buscador pertenecen siempre a los mismos países, y en proporciones parecidas. Esto permite mantener un contexto para las consultas que ellos formulan.

De acuerdo a los resultados obtenidos, podemos decir que en general las áreas tienen comportamientos parecidos en cada mes, aunque sus consultas vayan cambiando, ya sea por los acontecimientos que suceden en el país o en el mes en específico, o porque los usuarios van generando nuevos tipos consultas (con más palabras, lo que hace difícil que se vayan repitiendo a lo largo de los meses). Estos comportamientos están dados por curvas que siguen distribuciones de Zipf [11], donde son pocas las consultas que se repiten en cada mes.

Si bien las consultas para cada área están influenciadas por el ranking del buscador, en general son representativas de su área. En el caso de las áreas educación, gobierno y militar, las consultas se encontraban en el contexto de lo que significaban sus áreas. Para el caso de las áreas comercio y particular, si bien muchas de sus consultas son representativas de

productos y servicios específicos, existen consultas muy populares que debido a que no son situadas en un contexto pierden información acerca de la necesidad del usuario, como sexo, fotos, tierra o porno (y en general para algunas consultas compuestas de una sola palabra, que es polisémica). Para el área militar, encontramos que las consultas son específicas, las cuales describen y representan de mejor forma el área. Sin embargo el número que es posible encontrar es reducido.

Para cada una de las áreas de clasificación es posible determinar tipos de búsqueda, dependiendo de la representatividad y tipo de sus consultas. La necesidad de navegación está presente en todas las áreas, en cambio el de transacción sólo en algunas. En todas también se observa la búsqueda de información, donde la búsqueda indirecta sobre un tema en especial es el más común, así como el listado de opciones de sitios Web. De esta manera es posible caracterizar cada una de las áreas con respecto a los tipos de búsquedas que se hacen para cada una, y entender de mejor forma lo que la consulta realmente quiere decir.

Finalmente podemos concluir que a medida que la caracterización de estas áreas sea mucho más fina y las consultas más representativas, puede tener un uso en clustering de consultas o recomendación de consultas.

REFERENCIAS

[1] TodoCL: Localizado en <http://www.todocl.cl>, 2000.

[2] R. Baeza-Yates, C. Castillo; “Relating Web Structure and User Search Behavior” (extended poster), Proceeding International 10th WWW Conference, Hong-Kong, China, May 2001.

[3] R. Baeza-Yates; “Information Retrieval in the Web: beyond current search engines”, International Journal of Approximate Reasoning, 34:97-104, 2003.

[4] R. Cooley, B. Mobasher, J. Srivastava; “Web mining: information and pattern discovery on the world wide web”. ICTAI, 558-567, 1997.

[5] S. Chakrabarti; “Mining the Web: Discovering Knowledge from Hypertext Data”, Morgan Kauffmann, 2003.

[6] R. Baeza Yates, F. Saint-Jean, C. Castillo; “Web Dynamics, Structure and Page Ranking”, SPIRE 2002, Springer LNCS, Lisboa, Portugal, 117-130, 2002.

[7] R. Baeza Yates, B. Poblete, F. Saint-Jean; “Evolución de la Web Chilena 2001-2002”, Estudio Técnico, Centro de Investigación de la Web, CIW, 2003.

[8] NIC Chile: Localizado en <http://www.nic.cl>.

[9] A. Broder; “A Taxonomy of Web Search”, ACM SIGIR Forum, 2002.

[10] D. Rose, D. Levison; “Understanding User Goals in Web Search”, Proceedings of the 13th conference on World Wide Web, NY USA, 2004.

[11] G. Zipf; “Selective Studies and the Principle of Relative Frequency in Language”, Harvard University Press, Cambridge, MA, 1932.