Relating Web Structure and User Search Behavior

Ricardo Baeza-Yates *Carlos Castillo*
Depto. de Ciencias de la Computación, Universidad de Chile
Blanco Encalada 2120, Santiago 6511224, Chile
E-mail: {rbaeza,ccastill}@dcc.uchile.cl *

Abstract

In this paper we present some empirical results relating the Web structure and user search behavior for the Chilean (.cl) Web domain. Some of our results shed light on the correlation between Web connectivity, Web pages, and queries posed to a Chilean search engine.

1 Introduction

The Web became popular in less than ten years and has grown exponentially to an estimated number of pages of over two billion. Several studies have characterized the Web size, connectivity, dynamics, user search behavior, and languages, to mention a few. However, there is little information about how all those characteristics relate to each other, in particular, which are the main dependencies. In this paper we explore the relation between the structure of the Web and user search behavior using the Chilean Web. Although this is a small subset of the Web, is not a sample of the global Web as in most other studies. In fact, all the pages of a country are much more homogeneous, as they share a culture, are dominated by a single language, and most page visits have a common context. In summary, our subset is very close to a logical collection of pages.

In the first half of 2000, we collected about 670 thousand pages of the Chilean (.cl) domain, corresponding to approximately 7.500 Web sites. About 93% of the pages are in Spanish, while most of the rest are in English, with an average page size of about 15Kb. The .cl domain currently has about one million pages and more than 10 thousand sites and also grows exponentially, albeit perhaps slower than all the Web. Our data comes from the TodoCL search site [3] which specializes on the Chilean Web and is part of a family of vertical search engines built using the Akwan search engine [2]. TodoCL also has a directory which is based on the Open Directory Project [1], which at the time of the crawling had about three thousand entries for Chile. A complete characterization of the Chilean Web was presented in [4].

The most complete study of the Web structure [8] focus on page connectivity. One problem with this is that a page is not a logical unit (for example, a page can describe several documents and one document can be stored in several pages.) Hence, we decided to study the structure of how Web sites are connected, as Web sites are closer to be real logical units. Not surprisingly, we

*This work was partially supported by Fondecyt project 99-0627 and TodoCL.

found that the structure in Chile at the Web site level was similar to the global Web and then we use the same notation of [8], that is:

(a) MAIN, sites that are in the strong connected component of the connectivity graph of sites;

(b) IN, sites that can reach MAIN but cannot be reached from MAIN;

c) OUT, sites that can be reached from MAIN, but there is no path to go back to MAIN; and

d) other sites that can be reached from IN (t.in), sites in paths between IN and OUT (tunnel), sites that only reach OUT (t.out), and unconnected sites (island).

We extend this notation by dividing the MAIN component into four parts:

(a) MAIN-MAIN, which are sites that can be reached directly from the IN component and can reach directly the OUT component;

(b) MAIN-IN, which are sites that can be reached directly from the IN component but are not in MAIN-MAIN;

(c) MAIN-OUT, which are sites that can reach directly the OUT component, but are not in MAIN-MAIN;

(d) MAIN-NORM, which is the rest.

We found that this division of the MAIN component was interesting, but in other countries does not work as well (for example, in Spain).

Figure 4 shows the percentage of pages in each component (the left column), while figure 1 shows the structure using number of pages and number of Web sites of each component to represent the area of each part of the diagram.

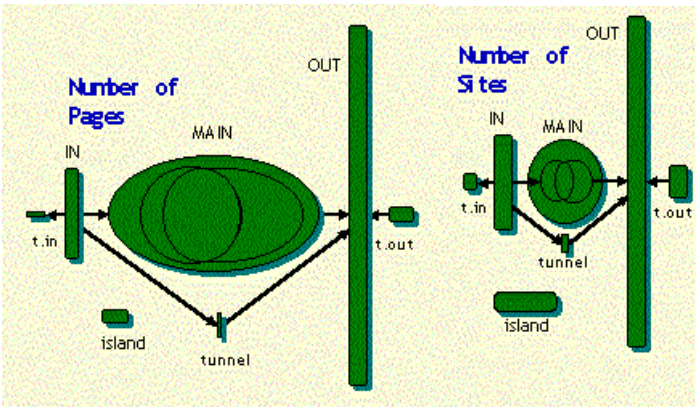

Figure 1: Connectivity structure of the Chilean Web with component areas proportional to number of pages, and number of sites.

2 Web Search Behavior

Search engines are one of the most visited Web sites and several studies show that most visits are the result of a Web search. The use of this type of tool depends on the user expertise [9]. As is customary, TodoCL keeps track of user behavior, and in particular, queries submitted to it. In this study, we used 730 thousand queries in a period of about three months. Those queries had an average length of 2.43 words (this is similar to the AltaVista study [10]), and 29% of the queries had at least one stopword in it (stopwords are words that are not useful in most cases for a search because they appear in almost all pages). The queries do not have operators because TodoCL uses a menu with three alternatives (search for all the words, some of the words, or a sentence).

The Web collection has approximately two million different words. It is well known that the size of the vocabulary follows a sub-linear model (Heaps' law [7]) with an exponent around .5 for English text data. In our Web collection the exponent goes up to .63, which is consistent with the fact that we have two languages, many more mistakes, and other sequences that are not words in any natural language. On the other hand, most of the words in the queries did appear in the collection.

One interesting related issue is how queries (which can be seen as small documents) differ from Web pages (document collection). Figure 2 show the word frequency distribution in the text collection and the queries as well as the document frequency distribution of the words in the collection.

It is well known that word frequency can be modeled by a generalized Zipf distribution, where the frequency of the i-th word is proportional to $i^{-\theta}$ (for example, see [7, ch. 6]). Our intuition was that query words were more biased than the words in the collection, because there are very popular terms such as MP3. Using least squares on the central part of our data (that is, eliminating most frequent words and the right tail) we obtained the following parameters for them: $\theta = 1.59$ for the collection (term or document frequency) and $\theta = 1.24$ for the queries. That is, queries are less skewed than words in the collection. Two of the models are also plotted in figure 2 (the third model is a line parallel to the bottom one). Another unexpected result is that the document and term distribution in the collection are almost parallel and they only meet for very infrequent words (instead of approaching each other slowly).

How the queries relate to the collection? Figure 3 shows the normalized frequency of the words in the queries using the frequency order of the words in the collection. The fact that queries are less skewed is corroborated by the green dots over the red line, which are more frequent on the right. On the other hand, stopwords in the queries appear below the red line (green region at the left).

To relate search behavior to the Web structure, we used the information of which pages were visited after a search. Figure 4 shows the fraction of sites in each component visited after a search with respect to Web structure as well as which pages are chosen by ODP editors to build the directory. We can clearly see that searching users choose pages very differently from ODP editors. One reason could be that the behaviors are similar, but the choices for good pages are not. Notice that because the ODP links are inside TodoCL, and TodoCL belongs to MAIN, there cannot be ODP pages in the IN component. To solve this problem, we excluded TodoCL from MAIN in this analysis.

Figure 4 shows that for the users, the Web structure is different than, say the collection itself or

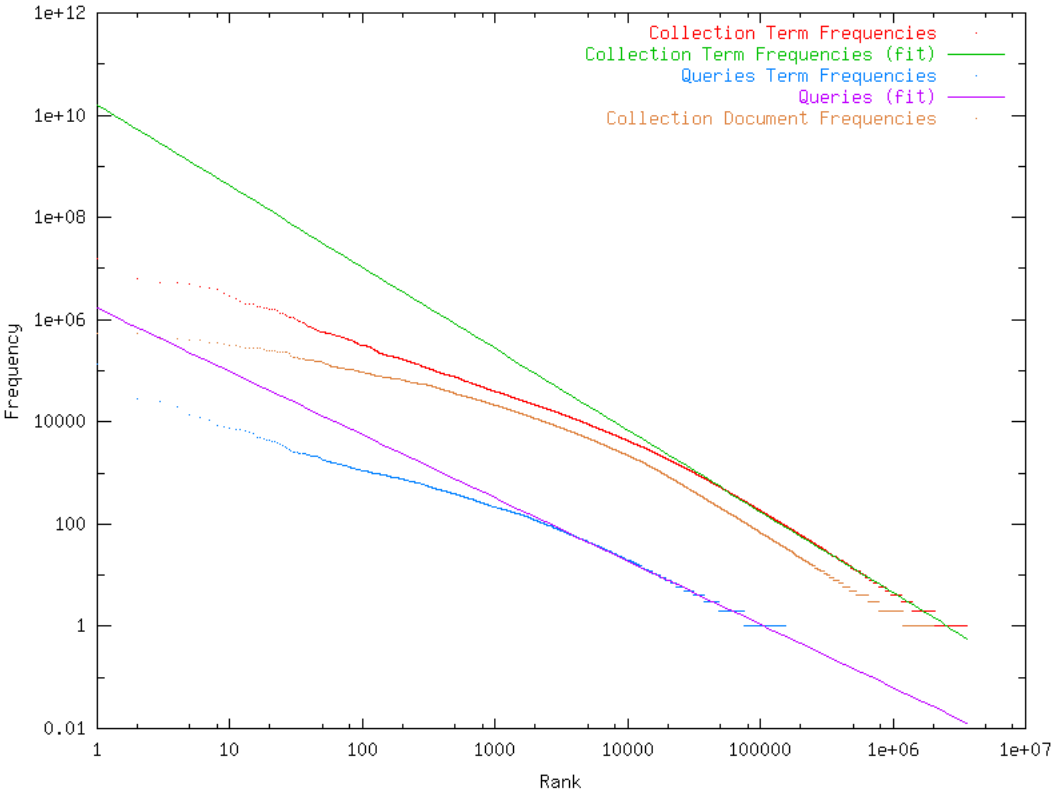

Figure 2: Term and document frequency in the collection and term frequency in the queries.

ODP editors (which have very restrictive policies). This means that the search engine is discriminating the sites, guiding the users to good resources. In fact, if the proportions were the same for the search engine and the directory, would mean that the search engine is biased to popular and old sites. This type of diagram can be used to evaluate a ranking algorithm and obtain a distribution accordingly to a certain goal. For example, more uniform or biased towards newer sites. Perhaps the most interesting relation is the bias of Web editors to older sites (possibly because are easier to find due to the bias of ranking algorithms of Web search engines), which affects the information of good directories.

3 Conclusions

In this paper we have attempted a first study to correlate Web characteristics and user search behavior. One first criticism might be the data size. Although one million pages is small nowadays, is big enough for a statistical study. In addition, we have the advantage that we can crawl .cl almost completely (over the 95% of the Web sites), which is not the case in larger studies, and is not biased to "popular" or "better" pages. That is, as the coverage is larger, the results are in some sense more complete.

We can argue that Web sites in the MAIN component are better on average than in other

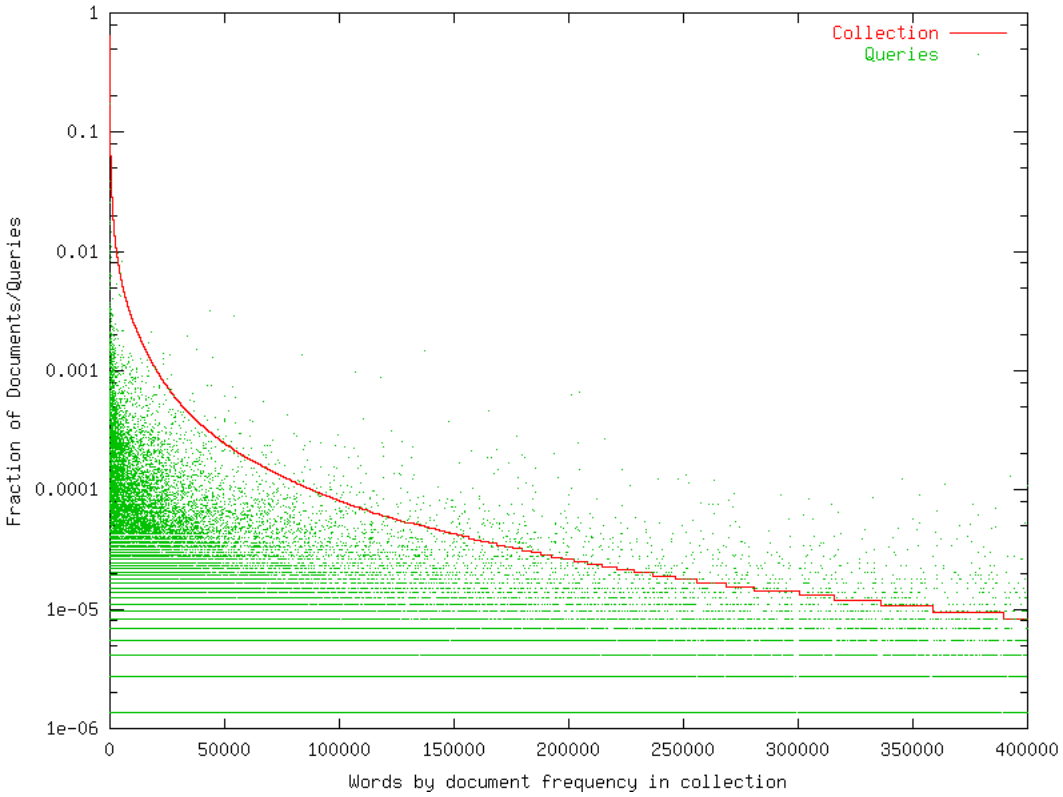

Figure 3: Relation between the words in the collection and the queries.

components (through link analysis), which might imply a correlation between retrieval quality and the Web structure. We have explored this by relating Web structure with link-based ranking algorithms [6]. Further results on relating Web characteristics and Web structure are available in [5].

References

[1] Open directory project: `http://odp.org`, 1999.

[2] Akwan: `http://www.akwan.com`, 2000.

[3] Todocl: `http://www.todocl.com`, 2000.

[4] R. Baeza-Yates and C. Castillo. Characterizing the Chilean Web (in Spanish). In *Chilean Computer Science Congress*, Santiago, Chile, Nov 2000.

[5] R. Baeza-Yates and C. Castillo. Relating Web characteristics. Technical report, CS Dept., Univ. of Chile, Santiago, Chile, Dec 2000. Available in www.ricardo.cl/ftp/relating.ps.gz.

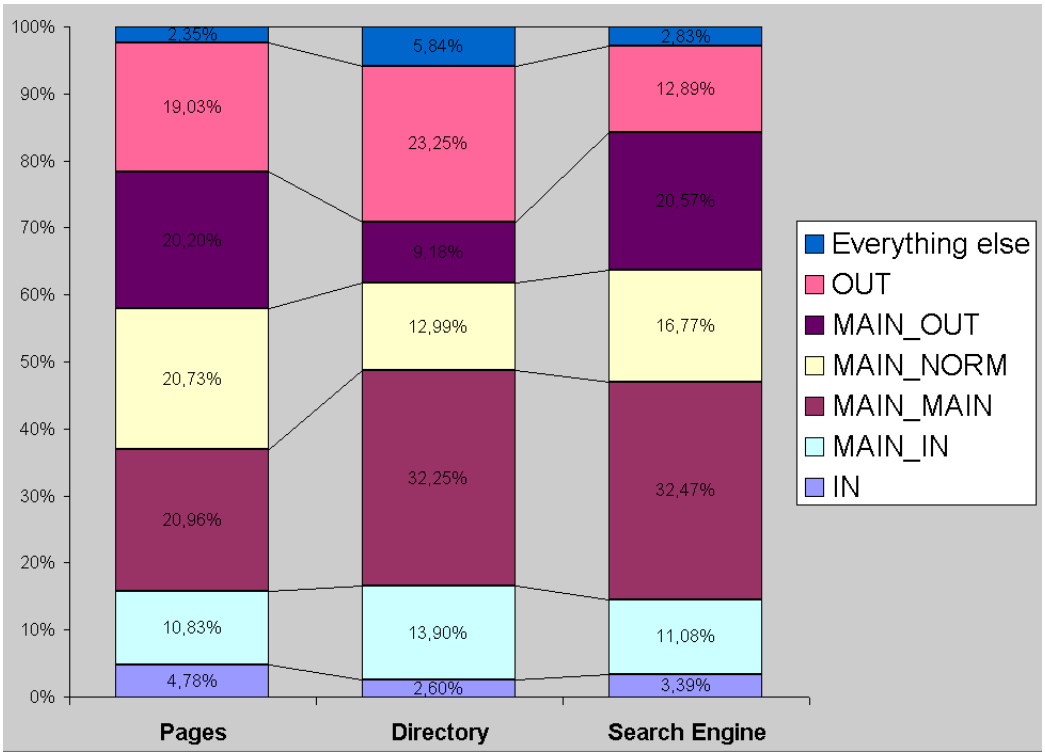

Figure 4: Percentages of pages chosen by the searchers and the ODP editors in the different components.

[6] R. Baeza-Yates and C. Castillo. Analysis of link based ranking for the Web. Technical report, CS Dept., Univ. of Chile, Santiago, Chile, Jan 2001. Available in www.ricardo.cl/ftp/linka.ps.gz.

[7] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison-Wesley & ACM Press, Harlow, UK, 1999.

[8] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web: Experiments and models. In *9th Int. WWW Conference*, Amsterdam, Holand, May 2000.

[9] C. Holscher and G. Strube. Web search behavior of internet experts and newbies. In *9th Int. WWW Conference*, Amsterdam, Holand, May 2000.

[10] C. Silverstein, M. Henzinger, J. Marais, and M. Moricz. Analysis of a very large Alta Vista query log. Technical Report 1998-014, Compaq Systems Research Center, Palo Alto, CA, USA, 1998.