Mining the Query Logs of a Chinese Web Search Engine for Character Usage Analysis

Yan Lu
School of Business
The University of Hong Kong
Pokfulam, Hong Kong
isabellu@business.hku.hk

Michael Chau
School of Business
The University of Hong Kong
Pokfulam, Hong Kong
mchau@business.hku.hk

Xiao Fang
College of Business Admin.
The University of Toledo,
Toledo, Ohio 43606, USA
xiao.fang@utoledo.edu

Abstract

The use of non-English Web search engines has been prevalent. Given the popularity of Chinese Web searching and the unique characteristics of Chinese language, it is imperative to conduct studies with focuses on the analysis of Chinese Web search queries. In this paper, we report our research on the character usage of Chinese search logs from Web search engine in Hong Kong. By examining the distribution of search queries terms, we found that people intended to use more diversified terms and that the usage of characters in search queries was quite different from the character usage of both general online information in Chinese and Web search queries in English. We believe the findings from this study have provided some insights into further research in non-English Web searching and will assist in the design of more effective Chinese Website search engines.

Keywords: web searching, character usage, search log analysis

1. Introduction

With the increasing popularity of the World Wide Web as a major information resource for people worldwide, research interests in analyzing search engine logs to understand Web users' search behavior has rapidly increased. Although many studies have been conducted on the query logs in search engines that are primarily English-based (e.g., Excite and AltaVista), the analysis of character usage in the Web search engines in non-English languages has been less investigated.

As the number of non-English resources increases dramatically on the Web, it is of great importance to study users' Web searching behavior for non-English contents using non-English search engines. Such research will provide important indications to improve the design of non-English search engines. Moreover, different languages have different characteristics. Previous findings of English search engines may not be applicable to non-English search engines. For example, different from English, Chinese is a character-based language. For Chinese, most of the meaningful words are built up by combining single characters together, and an individual word may not accurately indicate its real meaning in all the search queries it belongs to. Due to this specific characteristic of Chinese, traditional data processing methods for English search queries cannot be applied to process Chinese search queries. In this study, we choose character-based processing method for analyzing Chinese search queries, which has been widely used in the analysis of Chinese text (Chau et al., 2005b; Li et al., 2002).

In this paper, we report our analysis of the search query logs collected from a Chinese Website search engine called Timway. The rest of the paper is organized as follows. Previous studies on search log analysis and online Chinese are reviewed in Section 2. We propose our

research questions in Section 3. The data and the methods we used in this research are discussed in Section 4. The findings of our analysis are presented in Section 5. We conclude the paper in Section 6 with a summary of our study and some future directions.

2. Related Studies

2.1 Studies on Web Search Queries

Large-scale studies on general-purpose English search engines such as Excite and Alta Vista began at the end of 90's (Jansen et al., 1998; Jansen et al., 2000; Spink et al., 2001; Wolfram et al., 2001; Silverstein et al., 1999). Interesting findings of these studies include topic trends in Web searching (Spink et al., 2002), sex related information searching on the Web (Spink et al., 2004), and characteristics of question format Web queries (Spink & Ozmultu, 2002).

Another category of Web search analysis examined the search logs of a specific Website or an information system. Croft et al. (1995) analyzed search data obtained from THOMAS and found that 88% of all queries contained three or fewer words, which was much lower than the number of words contained in queries to a traditional information retrieval system. Jones et al. (1998) studied transaction logs of the New Zealand Digital Library and obtained similar results to those reported in Croft et al. (1995). They found that almost 82% of queries were composed of three or fewer words. Chau et al. (2005a) and Wang et al. (2003) studied search queries submitted to a search engine in a government website, and search queries submitted to a search engine in a university website respectively. Both of the studies found that information search behavior when using a general-purpose search engines was quite different from information search behavior when using a Website specific search engine.

Few studies have focused analyzing user information seeking behavior when using a non-English search engine. Hoelscher (1998) analyzed searching data from Fireball (http://www.fireball.de), a German Web Information Retrieve system. Data set used in the study contained about 16 million queries and 27 million non-unique terms. Some summary statistics, such as the average length of queries, the use of Boolean operators, and the use of phrase searching, were discussed in the paper. Huang et al. (2001; 2003) analyzed query logs from two Chinese search engines in Taiwan, namely GAIS (gais.cs.ccu.edu.tw) and Dreamer (no longer available). Instead of studying users' information needs and searching behaviors from the search logs, they utilized query logs to recommend term suggestions to users. They also proposed a method to extract search sessions and search queries from proxy server logs. They found that 74% of search sessions contained only one query, which was close to the number reported in the AltaVista study (Silverstein et al., 1999). Similar to the Fireball study, the major drawback of the Taiwan search engine study is that only limited statistics were provided; and there was no in-depth analysis of query terms and search topics.

Previous studies on web search queries used a set of similar statistics, focusing on three different levels-sessions, queries and terms (Spink et al., 2001; 2002; Silverstein et al., 1999; Jansen et al. 1998; Jansen et al. 2000). These statistics allow researchers to compare their findings across different types of search engines at different times.

3. Research Questions

As discussed above, few studies have focused on character usage in Chinese search engines. Most of the previous Web searching studies focused on English search engines. Although some of them analyzed the characteristics of search terms, their findings may not be applicable to Chinese search engines, due to the great discrepancy between Chinese and English. The purpose of our study was to explore the character usage of search terms

submitted by Chinese search engine users, and to assist in the design of more effective Chinese Website search engines.

We sought to answer the following research question in our study:
- (1) What are the characteristics of the character usage of the search queries submitted to Chinese Website search engines?
- (2) How do these character usages compare to those of Chinese general online texts as well as those of English search engine like Excite?

4. Data and Methods
We collected query logs of the Timway Search Engine (http://www.timway.com). Timway, a Chinese search engine established in 1997, is primarily designed for searching Web sites in Hong Kong. It supports search queries in both Chinese and English, and indexes Web pages in both languages.

The query log used in our study covers a three-month time period from December 1, 2003 to March 2, 2004. It consists of 1,255,633 records in total. Each record represents a search query submitted to the search engine. One record consists of four fields: search query, number of hits, user's IP address, and timestamp (Chau et al., forthcoming).

Out of the 1,255,633 queries, 536,814 are Chinese queries, 641,169 are English queries, and 77,650 are mixed queries. We focus on analyzing Chinese queries in this study. As most of the Chinese queries in our data are originally in Big 5, we used an open-source Java program to convert all queries in GB-2313 and GBK into Big 5.

5. Analysis Results
This section begins by calculating the number of Chinese characters in one query in comparison with the number of English words reported in Excite by Spink et al. (2001). Next is the comparison between the character usage of search queries from Timway and that of another two Chinese corpora as well as an English search web corpus. This is followed by an analysis of the distribution of n-grams (n=1 to 6) extracted from Chinese search logs.

5.1 Number of Tokens per Query
Mean number of characters in Chinese queries is 3.380, which is much larger than the mean number of words in English queries as reported in Excite (2.16) by Spink et al. (2001).
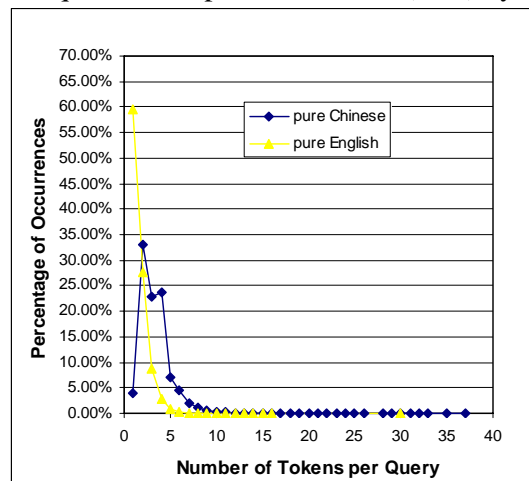

Figure 1: Comparison of Number of Tokens per Query

Figure 1 compares between the number of tokens per query for pure Chinese corpus and that for pure English corpus. Search queries in both languages are generally short. The curve for pure Chinese queries fluctuates with two crests, which represent two and four characters per query respectively. This indicates that most pure Chinese queries consist of two, three, or four characters. The main reason for this phenomenon lies in the characteristics of Chinese language. Each single Chinese character corresponds to a morpheme, the smallest meaningful linguistic unit that cannot be divided into smaller meaningful parts. Although some meaningful words consist of one character, most of the meaningful words in Chinese consist of at least two characters. As shown in Figure 1, bigrams, trigrams, and 4-grams are most frequently used search terms in Chinese.

5.2 Character Usage of Chinese Search Queries

5.2.1 Comparison with Chinese corpora

To study character usage of Chinese search queries, we ranked the search log queries and compared top 50 of them with those of the MTSU corpus (Da, 2004), and those of the Usenet Newsgroups corpus (Tsai, 1996).
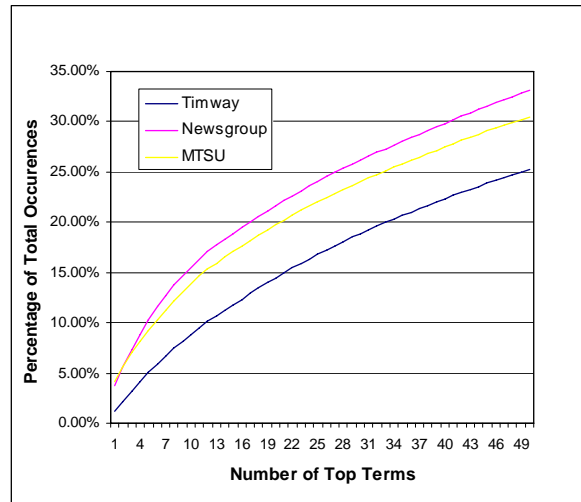

Figure 2: Comparison of Top 50 Terms Distribution with the Timway Search Log Data, the Usenet Newsgroup Corpus, and the MTSU Corpus

Figure 2 illustrates the distribution of top 50 unigram characters from Timway, newsgroup, and MTSU. The comparison with these three curves shows that the percentile of total occurrences of Newsgroup grows at a slightly faster rate than that of MTSU and at an obviously faster rate than that of Timway. The curve of Timway has the lowest growing rate among them, indicating that web site query terms contain a very large number of different terms compared with large Chinese texts in general, even though the general Chinese texts are collected from online resources.

We identified the overlapping unigrams and bigrams between Timway corpus and MTSU corpus separately in the range of the top 5,000 words for each corpus. Figure 3 presents the result.
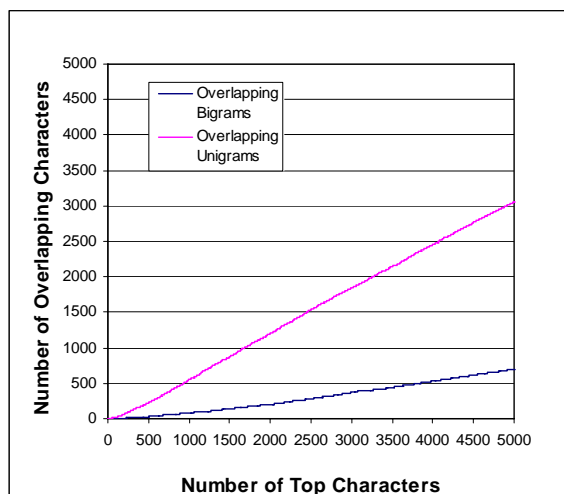
Figure 3: Overlapping Unigrams, Bigrams between Timway and MTSU

For unigram characters, these two corpora have 3,056 overlapping characters, 61.12% out of 5,000 single characters. The curve for overlapping bigrams increases much more gently than the one for overlapping unigrams. Only 698 bigrams out of the top 5,000 bigrams from Timway appear in the top 5,000 bigrams from MTSU, with the percentage of 13.96%. The great discrepancy between these two curves implies that although the two corpora have quite a lot of same unigrams, they have few in common for bigrams which are more topic-related.

Among the top 100 most frequently occurring bigrams of the two corpora, there are only two overlapping words - "中國" (China) coming at the 9th place of the Timway bigram list and at the 60th rank of MTSU list, and "電話" (telephone) appearing at the 28th of the Timway list and at the 78th of the MTSU list. The extremely limited overlap of these two corpora further verifies that the language of search engine queries has its unique characteristics, especially for Chinese search engines, which accept large amount of queries in Chinese. This warrants further study of the character usage of Chinese search queries.

5.2.2 Comparison with English corpus

To further our analysis, we compare the distribution of top 50 unigrams of Timway search queries with that of Excite. The Excite data used for comparison is reported in (Spink et al, 2001). The data they analyzed consisted of a log of transaction record of 1,025,910 queries submitted by 211,063 users on 16 September 1997. There were 1,277,763 search terms in all unique queries.

As shown in Figure 4, the top 50 unigram queries from Timway represented 25.25% (510,570) of all the unigram queries, approximately 2.5 times greater than the Excite data, in which the top 50 unigram queries represented only 10.15% of all its unigram queries. (The percentage we got here is different from the percentage reported by Spink et al. 2001, because they use unique queries for calculation while we use all the queries in this study.) This visible discrepancy is highly due to the limited number of Chinese characters. We found that out of the 536,814 Chinese queries in our data, there are only 7,303 unique Chinese characters. This finding is very different from the Excite corpus which has 140,279 unique English terms out of approximately 1 million queries in total (Spink et al., 2001).
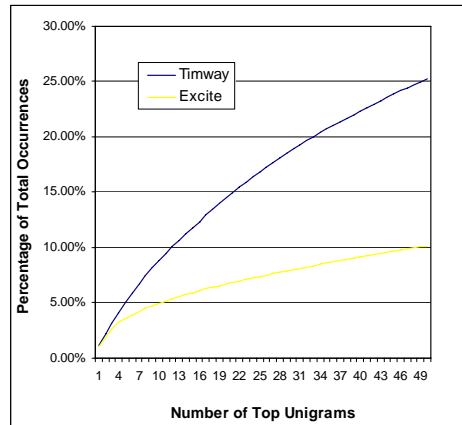
Figure 4: Comparison of Accumulation Curves for Top 50 Unigrams with the Timway
Search Log Data and the Excite Search Log Data

6. Discussions

This article is among few studies investigating the characteristics of Web searching in Chinese. Some valuable findings on non-English information search behavior are revealed in this article.

By comparing the character usage of our corpus with that of two corpora obtained from online general texts, we found that our data had the lowest growing rate for the distribution of occurrence. We speculated that people intended to use more diversified terms when searching online. The little overlap in high-frequency characters between our corpus and each of the two corpora further confirmed that the usage of characters for search queries was quite different from that for general online information.

The research also identified the discrepancy between the distribution of our corpus and that of the Excite corpus (Spink et al., 2001). The visible high degree of usage of the most frequent terms in Timway corpus with the percentage of 25.25%, compared with relatively low percentage of 10.15% in Excite corpus, provides an implication for search engine designers that the classification of website content and the index items of quick searching could be organized in a more precise and easy-using way.

Generally, the language of Chinese Web queries has its own and unique characteristics, it is more advisable for search engine designers to extract potential search topics by analyzing the search terms submitted by users rather than analyze the available texts online.

7. Conclusion and Future Directions

In this paper, we report our study on the character usage of a Chinese Website search engine. Our research has identified several characteristics of search terms of our corpus by comparing with other corpus from different perspectives. Due to the lack of studies on the search logs in Chinese, we were unable to undertake comparison with analogous corpora, thus limited our research scope. More studies on the search logs in Chinese and other non-English languages are highly desired.

References

Chau, M., Fang, X., and Liu Sheng, O. R., "Analysis of the Query Logs of a Web Site Search Engine", *Journal of the American Society for Information Science and Technology*, (56: 13), 2005a, pp. 1363-1376.

Chau, M., Qin, J., Zhou, Y., Tseng, C., and Chen, H., "SpidersRUs: Automated Development of Vertical Search Engines in Different Domains and Languages", in *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, Denver, Colorado, USA, June 7-11, 2005b.

Chau, M., Fang, X., and Yang, C., "Web Searching in Chinese: A Study of a Search Engine in Hong Kong, *Journal of the American Society for Information Science and Technology*, forthcoming.

Croft, W. B., Cook, R., and Wilder, D., "Providing Government Information on the Internet: Experiences with THOMAS", in *Proceedings of the Digital Libraries '95 Conference*, Austin, Texas, 2004, pp. 19-24.

Da, J., Chinese Text Computing. [Online], retrieved from: http://lingua.mtsu.edu/chinese-computing on Nov 05, 2005.

Hölscher, C., "How Internet Experts Search for Information on the Web", in *Proceedings of the World Conference of the World Wide Web, Internet, and Intranet*, Orlando, Florida, USA, 1998.

Huang, C. K., Chien, L. F., and Oyang, Y. J., "Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs", *Journal of the American Society of Information Science and Technology*, (54:7), 2003, pp. 638-649.

Huang, C. K., Oyang, Y. J., and Chien, L. F., "A Contextual Term Suggestion Mechanism for Interactive Search", in *Proceedings of The First Web Intelligence Conference (WI'2001)*, Japan, 2001, pp. 272-281.

Jansen, B. J. and Pooch, U., "Web User Studies: A Review and Framework for Future Work", *Journal of the American Society of Information Science and Technology*, (52:3), 2000, pp. 235-246.

Jansen, B. J., Spink, A., Bateman, J., and Saracevic, T., "Real Life Information Retrieval: A Study of User Queries on the Web", *ACM SIGIR Forum*, (32:1), 1998, pp. 5-17.

Jansen, B. J., Spink, A., and Saracevic, T., "Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web", *Information Processing and Management*, (36), 2000, pp. 207-227.

Jones, S., Cunningham, S. J., and McNam, R., "Usage Analysis of a Digital Library", in *Proceedings of the Third ACM Conference on Digital Libraries*, Pittsburgh, PA, USA, June 1998, pp. 293-294.

Li, Y., Ding, X., and Tan, C. L., "Combining Character-based Bigrams with Word-based Bigrams in Contextual Postprocessing for Chinese Script Recognition", *ACM Transactions on Asian Language Information Processing*, (1:4), 2002, pp. 297-309.

Silverstein, C., Henzinger, M., Marais, H. and Moricz, M., "Analysis of a Very Large Web Search Engine Query Log", *ACM SIGIR Forum*, (33:1), 1999, pp. 6-12.

Spink, A., and Ozmultu, H. C., "Characteristics of Question Format Web Queries: An Exploratory Study", *Information Processing and Management*, (38), 2002, pp. 453-471.

Spink, A., Ozmutlu, H. C., and Lorence, D. P., "Web Searching for Sexual Information: An Exploratory Study", *Information Processing and Management*, (40), 2004, pp. 113-123.

Spink, A., Wolfram, D., Jansen, B. J., and Saracevic, T., "Searching the Web: The Public and Their Queries", *Journal of the American Society for Information Science and Technology*, (52:3), 2001, pp. 226-234.

Tsai, C.-H., "Frequency and Stroke Counts of Chinese Characters." [Online], retrieved from: http://technology.chtsai.org/charfreq/ on May 20, 2005.

Wang, P., Berry, M. W., and Yang, Y., "Mining Longitudinal Web Queries: Trends and Patterns", *Journal of the American Society for Information Science and Technology*, (54:8), 2003, pp.743-758.

Wolfram, D., Spink, A., Jansen, B. J., and Saracevic, T., "Vox Populi: The Public Searching of the Web", *Journal of the American Society for Information Science and Technology*, (52:12), 2001, pp. 1073-1074.