Information Processing and Management 45 (2009) 115–130

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Characteristics of character usage in Chinese Web searching

Michael Chau[a,*], Yan Lu[a], Xiao Fang[b], Christopher C. Yang[c]

[a] *School of Business, The University of Hong Kong, Pokfulam, Hong Kong*
[b] *College of Business Administration, The University of Toledo, Toledo, OH 43606, USA*
[c] *College of Information Science and Technology, Drexel University, Philadelphia, PA 19104, USA*

ARTICLE INFO

Article history:
Received 4 July 2007
Received in revised form 28 April 2008
Accepted 10 June 2008
Available online 8 August 2008

Keywords:
Web mining
Search log analysis
Character usage
Chinese
Zipf distribution

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 a Web search engine in Hong Kong. By examining the distribution of search query terms, we found that users tended to use more diversified terms and that the usage of characters in search queries was quite different from the character usage of general online information in Chinese. After studying the Zipf distribution of n-grams with different values of n, we found that the curve of unigram is the most curved one of all while the bigram curve follows the Zipf distribution best, and that the curves of n-grams with larger n ($n = 3–6$) had similar structures with β-values in the range of 0.66–0.86. The distribution of combined n-grams was also studied. All the analyses are performed on the data both before and after the removal of function terms and incomplete terms and similar findings are revealed. 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 Web search engines.

1. Introduction

With the increasing popularity of the World Wide Web as the major information resource for people worldwide, research interests in analyzing search engine logs to understand Web users' search behavior have rapidly increased. A large number of studies have been conducted on the query logs in search engines that are primarily English-based (e.g., Excite and AltaVista); the query logs in the Web search engines in non-English languages have been less investigated. As the number of non-English resources on the Web increases considerably, 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 for improving 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, languages such as Chinese are character-based rather than word-based. In these languages, most of the meaningful words are built up by combining single characters together, and an individual character may not accurately indicate its real meaning in all the search queries it belongs to. For instance, the character "手" (hand) in Chinese can be combined with other characters to have different meanings, such as

[*] Corresponding author. Tel.: +852 28591014.
E-mail addresses: mchau@business.hku.hk (M. Chau), isabellu@business.hku.hk (Y. Lu), xiao.fang@utoledo.edu (X. Fang), chris.yang@ischool.drexel.edu (C.C. Yang).

0306-4573/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2008.06.005

"手機" (mobile phone), "二手" (second hand) and "手槍" (gun). Due to the specific characteristics of Chinese, traditional data processing methods for English search queries cannot be directly applied to the processing of Chinese search queries. We believe that it is interesting and timely to study the characteristics of the queries in non-English search engines.

In this paper, we report our analysis of the search query logs collected from a Chinese Web search engine called Timway (http://www.timway.com/). In particular, we study the characteristics on the usage of Chinese characters in their queries. We believe our findings are valuable for both relevant future research and the designers of Chinese Web search engines. The rest of the paper is organized as follows. Previous studies on search log analysis, online Chinese texts, and the application of Zipf distribution on search term analysis are reviewed in Section 2. We pose 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 a discussion on future directions and limitations.

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 the 90's (Jansen, Spink, Bateman, & Saracevic, 1998; Jansen, Spink, & Saracevic, 2000; Silverstein, Henzinger, Marais, & Moricz, 1999; Spink, Jansen, Wolfram, & Saracevic, 2002; Spink, Wolfram, Jansen, & Saracevic, 2001; Wolfram, Spink, Jansen, & Saracevic, 2001). Interesting findings of these studies include topic trends in Web searching (Spink et al., 2002), sex related information searching on the Web (Spink, Ozmutlu, & Lorence, 2004), and characteristics of question format Web queries (Spink & Ozmultu, 2002). In general, most studies have found that information seeking behaviors on the Web are quite different from that in traditional online information systems and suggested that more research in this area be necessary.

A few studies have focused on the analysis of users' information seeking behavior in non-English search engines. One example is the Fireball study (Hölscher, 1998). Fireball is a Web search engine in German (http://www.fireball.de). In that study, a dataset containing about 16 million queries and 27 million non-unique terms was analyzed. 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. The limitations of this study, as described by Jansen and Pooch (2000), include that no information was provided concerning user sessions, that discussion of query terms was limited, and that little descriptive information about the Fireball search engine was provided. Huang, Oyang, and Chien (2001) Huang, Chien, and Oyang (2003) analyzed the query logs from several Chinese search engines in Taiwan. Instead of studying users' information needs and searching behaviors from the search logs, they utilized query logs to provide term suggestions to users. They also proposed a method to extract search sessions and search queries from proxy server logs. Query logs used in the study contained search requests submitted to several general-purpose Web search engines in Taiwan, including GAIS (http://gais.cs.ccu.edu.tw), Dreamer (no longer available), Yahoo-Taiwan (http://tw.yahoo.com), Sina-Taiwan (http://www.sina.com.tw), PChome (http://www.pchome.com.tw) and Yam (http://www.yam.com). 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 studies was that only limited statistics were provided; and there was no in-depth analysis of query terms and search topics. Pu, Chuang, and Yang (2002) conducted another analysis of Chinese search logs collected from three search engines in Taiwan, namely Dreamer, GAIS, and Openfind (http://www.openfind.com.tw). They reported that average length of Chinese queries was 3.18 characters and that less than 5% of queries represented almost three-quarters of the total frequencies. They also noticed that users seldom used advanced search functions during their searching procedures (Pu et al., 2002).

Another category of Web search analysis examined the search logs of a specific Web site or an information system. Croft, Cook, and Wilder (1995) analyzed search data obtained from THOMAS, an online searchable database consisting of US legislative information. They 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, Cunningham, and McNam (1998) studied the transaction logs of the New Zealand Digital Library and obtained result similar to that reported in Croft et al. (1995). They found that almost 82% of queries were composed of three or fewer words. Chau, Fang, and Liu Sheng (2005) Wang, Berry, and Yang (2003) studied search queries submitted to a search engine in a government Web site and search queries submitted to a search engine in a university Web site, respectively. Both of the studies found that users' information search behavior in general-purpose search engines was quite different from that in a Web site specific search engine.

2.2. Search term usage analysis

Previous studies on Web search queries used a set of similar statistics in their studies, focusing on three different levels - sessions, queries and terms (Chau et al., 2005; Chau, Qin, Zhou, Tseng, & Chen, 2005; Jansen et al., 1998; Jansen et al., 2000; Silverstein et al., 1999; Spink et al., 2001, 2002). Statistics at the session level include the number of sessions, the number of queries in a session, changes in queries during a session, number of logic and modifiers, the number of result pages viewed by each user, and the use of relevance feedback. Statistics at the query level include the number of queries, the number of queries per user, the number of search terms in a query, the use of logic and modifiers, and the percentage of queries using

Boolean queries. Statistics at the term level include term rank and frequency distribution and the most highly used search terms. These statistics allow researchers to compare their findings across different types of search engines at different times.

Many studies applied the Zipf distribution to analyze the distribution of search terms. Zipf distribution, traditionally often applied to extensive textual passages, has been investigated for database contents in bibliographic and full text databases (Zipf, 1949). Let f be the frequency of a word in a corpus and γ be the rank of the word. According to the Zipf distribution,

$$f = k/\gamma \tag{1}$$

where k is a constant for a corpus. Zipf curves follow a straight line when plotted on a double-logarithmic diagram, which means when log(f) is drawn against log(r) in a graph, a straight line is obtained with a slope of -1. A number of theoretical developments of Zipf's law were later derived (Fedorowicz, 1982). A more general form of the Zipf distribution is as follows (Mandelbrot, 1953):

$$f = k/(\gamma + \alpha)^{\beta} \tag{2}$$

where α and β are constants for a corpus being analyzed. Generally, the constants α and β were found to have only small statistical deviations from the original law by Zipf (Smith & Devine, 1985).

Earlier Web search analysis studies have suggested that the distribution of terms used in Web search engines largely followed the Zipf distribution. Spink et al. (2001) used a double-log rank-frequency plot to determine the accordance of the Excite search log data with a Zipf distribution. The study found that the resulting distribution is slightly unbalanced for the high and low ranking terms. The findings concur with those of other previous studies (Nelson, 1989; Wolfram, 1992). Jansen et al. (2000) also found that the resulting distribution seemed to be unbalanced at the ends of the graph of rank-frequency distribution. The curve fell off very gently at the beginning and showed discontinuities and an unusually long tail toward the end. Wang et al. (2003) used Zipf's curve to analyze the usage of English search terms. Different from other studies, they drew a second line by ranking words based on unique frequencies and compared with the original line. The drastic drop of this second line indicated that the number of words with low frequency increased as the frequency decreased. Ha, Sicilia-Garcia, Ming, and Smith (2002) explored the validity of Zipf's law for large corpora in two languages, namely English and Chinese. They concluded with a confirmation of Zipf's original law in an extended form.

3. Research questions

Although some previous Web searching studies have analyzed the characteristics of search terms in English, their findings may not be applicable to Chinese search engines, due to the great discrepancy between the two languages. Only a few studies have focused on Web search logs in Chinese, and no previous research has explicitly studied the characteristics of character usage in Chinese search engines in depth. The purpose of our study is to explore the character usage of search queries submitted by Chinese search engine users. Analysis of the character usage can help us better understand the information seeking behavior in Chinese Web searching as well as provide insights to the design of more effective Chinese Web search engines.

We seek to answer the following research questions in our study:

(1) What are the characteristics of character usage in the search queries submitted to Chinese Web search engines?
(2) How do these characteristics compare to those of other online Chinese corpora?

4. Data and methods

We collected the query logs of the Timway Search Engine (http://www.timway.com) in our previous studies (Chau, Fang, & Yang, 2007). 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. *Search query* is the original search text entered by a user. *Number of hits* records the number of relevant search results found in Timway's database and returned to users. *User's IP address* and *timestamp* record where and when a query is submitted.

Out of the 1,255,633 queries, 536,814 are Chinese queries, 641,169 are English queries, and 77,650 are mixed queries. In this study, we focus on analyzing the Chinese queries only. 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.

In our previous study (Chau et al., 2007), we found that pornographic materials were the most sought-after in Chinese search engines and that the mean number and median number of queries per session in our search log were 2.03 and 1, respectively. We also found the mean number of characters used in Chinese queries to be 3.380, which was significantly larger than the mean number of terms in English queries reported in Excite (2.16) and AltaVista (2.35). The queries contain 7,303 unique Chinese characters, which was much lower than the number of unique terms in English queries. In addition, we found that the top 50 Chinese characters already represent one-quarter of all the Chinese characters in the search log.

This proportion was much higher than that in the Excite search log. All these findings suggested that Chinese search queries are quite different from English queries in many aspects. In the following, we focus our analysis on the characteristics of character usage in Chinese queries, which was not addressed in our previous study.

5. Analysis results

This section begins by analyzing the number of tokens (characters or words) per query in pure Chinese queries and in pure English queries submitted to Timway. Next we compare the character usage in the search queries of Timway and that of two other online Chinese corpora. This is followed by an analysis of the distribution of n-grams ($n = 1$–6) extracted from the Chinese search logs. An investigation on the search topics in which users are interested is discussed at the end of this session.

5.1. Number of characters per query

The mean number of characters in Chinese queries in our data is 3.380, which is slightly larger than the number of characters in Chinese queries reported by Pu et al. (2002) in their study on the Chinese search engine logs in Taiwan. However, our number is much larger than the mean number of words in English queries as reported in Excite (2.16) by Spink et al. (2001). Although Chinese characters are not linguistically equivalent to English words, they are compared here because both of them are recognized as the major indexing units in search engines.

Fig. 1 compares between the number of characters/words (we call them "tokens" here) per query in pure Chinese queries and that in pure English queries. Most search queries are short in both languages. The curve for pure Chinese queries fluctuates with two crests, which represent two and four characters per query, respectively. The curve shows that most pure Chinese queries consist of two, three, or four characters. The main reason for this phenomenon lies in the characteristics of the Chinese language. Each single Chinese character corresponds to a morpheme — the smallest meaningful linguistic unit that cannot be divided into smaller meaningful parts, and most of the commonly used "content words" in Chinese consist of two or more characters.

5.2. Comparison with Chinese corpora

5.2.1. Unigrams

To study the characteristics of character usage in Chinese search queries, we ranked the search terms and compared the top 50 frequently used terms with those of two online Chinese text corpora, namely the MTSU corpus (Da, 2004) and the Usenet newsgroups corpus (Tsai, 1996). The MTSU corpus has 9,933 unique characters out of 193,504,018 characters in total. The data include character frequency lists generated from a large collection of Chinese texts obtained from online sources. The data also include bigram frequency lists as well as individual mutual information scores generated from two sub-corpora. The other corpus we used, the Usenet newsgroups corpus, consists of all the BIG-5 Chinese characters that appeared in Usenet newsgroups and contains 171,882,493 characters as well as their frequencies (Tsai, 1996).

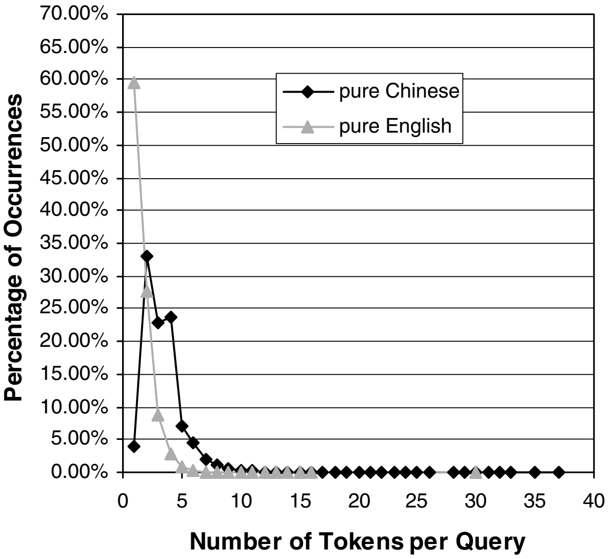

Fig. 1. Comparison of number of tokens per query.

Table 1a
The Top 25 Characters in the Timway search log data, the Usenet newsgroup corpus, and the MTSU corpus

Timway		Usenet newsgroup		MTSU	
Character	Percentage (%)	Character	Percentage (%)	Character	Percentage (%)
人[a,b]	1.19	的	3.80	的	4.09
中[a,b]	0.98	是	1.86	一	1.58
港	0.98	不	1.65	是	1.35
香	0.94	我	1.50	不	1.16
電	0.92	一	1.48	了	1.10
情	0.87	有	1.33	在	1.04
成	0.83	大[a]	1.10	人[b]	0.96
色	0.77	在	1.00	有	0.92
學	0.72	人[a]	0.93	我	0.88
小	0.66	了	0.88	他	0.82
會[a]	0.63	中[a]	0.77	這	0.80
圖	0.60	到	0.76	個	0.62
文	0.59	資	0.65	們	0.61
美	0.56	要	0.60	中[b]	0.57
大[a,b]	0.56	以	0.58	來	0.56
國[b]	0.56	可	0.58	上	0.55
手	0.55	這	0.57	大[b]	0.54
機	0.54	個	0.54	爲	0.54
女	0.53	你	0.53	和	0.52
樓	0.52	會[a]	0.52	國[b]	0.51
日	0.49	好	0.50	地	0.50
網	0.48	爲	0.49	到	0.50
天	0.46	上	0.48	以	0.47
生	0.45	來	0.47	說	0.46
子	0.44	學	0.47	時	0.43

[a] These characters are found in both the Timway data and the Usenet newsgroup corpus.
[b] These characters are found in both the Timway data and the MTSU corpus.

Table 1a lists the top 25 characters that appeared in these three corpora. Among the three listings shown in Table 1a, the Timway data have only four overlapping characters with the Usenet newsgroup data as well as the MTSU data. Reasons for few overlapping come from differences between the character usages in search queries and that in general texts on the Internet. In a corpus consisting of general texts, the most frequently used words are mostly function words that do not have much semantics and are less frequently used in search queries.

To make further comparison of character usage between search queries and online texts, we removed all the function words (i.e. stop words in Chinese) from these three corpora and do the comparison again. Table 1b lists the top 25 characters that appeared in these three corpora with function words removed.

Although the number of overlapping characters increases from four to ten as compared with the Usenet newsgroup corpus and from four to eight as compared with the MTSU Corpus, the overlapping rates still remain at a low level.

Fig. 2a illustrates the distribution of the cumulative occurrences of the top 50 unigram characters from Timway, the newsgroup corpus, and the MTSU corpus. The number of occurrences of a unigram is measured by the number of times that it appears in the search logs. Comparison among these three curves shows that the percentile of total occurrences in the newsgroup corpus grows at a slightly faster rate ($\beta = 0.5199$ for linear regression) than that of MTSU (0.4829), and at an obviously faster rate than that of Timway (0.4633).

Fig. 2b illustrates this distribution after removing function words and the percentage of total occurrences was calculated based on the occurrences of content-related unigrams only. After removing the function words, the slope of the Timway data only changes slightly from 0.4633 to 0.4855, while the slopes of Usenet newsgroup and MTSU decrease dramatically to 0.4505 and 0.3777, respectively. As shown in Fig. 2b, after removing function words, the curve of the Timway data and the newsgroup data almost grow at the same rate for the top 50 unigrams. However, the slope for the MTSU data grows at a much slower rate than the other two corpora.

The differences in Figs. 2a and 2b confirmed that Web query terms contain more content words and fewer function words than general online texts and newsgroup texts. When function words are included, unigrams in Web query are more diversified than general Chinese texts. When function words are excluded, in Web query can be less diversified. These differences

Table 1b
The top 25 characters in the Timway search log data, the Usenet newsgroup corpus, and the MTSU corpus with function words removed

Timway		Usenet Newsgroup		MTSU	
Character	Percentage (%)	Character	Percentage (%)	Character	Percentage (%)
人[a,b]	1.27	大[a]	1.73	人[b]	1.50
中[a,b]	1.05	人[a]	1.47	中[b]	0.88
港	1.04	中[a]	1.21	大[b]	0.84
香	1.00	資	1.02	國[b]	0.79
電[a]	0.98	會	0.82	說	0.70
情	0.92	好	0.79	時	0.67
成[b]	0.88	學[a]	0.74	會	0.59
色	0.82	交	0.67	生[b]	0.55
學[a,b]	0.77	時	0.60	子[b]	0.51
小[a]	0.70	文	0.60	年	0.48
會	0.67	說	0.59	發	0.46
圖	0.64	看	0.58	作	0.43
文	0.63	問	0.55	裏	0.43
美	0.59	生[a]	0.55	道	0.43
大[a,b]	0.59	提	0.55	行	0.43
國[b]	0.59	請	0.53	然	0.41
手	0.58	天[a]	0.52	家	0.41
機[a]	0.58	小[a]	0.49	種	0.40
女	0.56	想	0.48	事	0.40
樓	0.55	工	0.48	成[b]	0.40
日	0.52	還	0.47	方	0.39
網	0.51	電[a]	0.47	法	0.37
天[a]	0.49	機[a]	0.43	學[b]	0.37
生[a,b]	0.48	子[a]	0.42	現	0.35
子[a,b]	0.47	訊	0.41	動	0.34

[a] These characters are found in both the Timway data and the Usenet newsgroup corpus.
[b] These characters are found in both the Timway data and the MTSU corpus.

discussed above indicate that the study on Chinese search terms is imperative. The results from analyzing online Chinese corpus cannot be directly applied to studies in Chinese search engines.

5.2.2. Bigrams

Following the analysis of unigrams, we analyze the characteristics of bigrams in search queries. We conduct the comparison for bigrams between the search logs from Timway and the text from the MTSU corpus.[1] Table 2a shows the top 25 most frequently used bigrams from the Timway search logs and the MTSU dataset, respectively.

In Table 2a, we can see that the Timway data contain mostly content words such as "香港" (Hong Kong), "成人" (adult), and "色情" (pornography) while the MTSU data contain mostly function words or other non-content bearing phrases such as "一個" (one), "什麼" (what), and "沒有" (haven't). There is no overlap in the two lists. Similar to unigrams, we see that content words dominate Web queries and function words dominate general online texts, which is an expected observation.

To compare the topics in the two corpora, we need to remove the incomplete terms and function terms from the data. A bigram can be an incomplete or invalid term if it is a part of a longer term or parts of two terms. For example, a query "香港電影" (Hong Kong movies) will result in three raw bigrams: "香港", "港電", and "電影", where the second bigram is invalid. In Table 2a, the bigrams "六合" and "合彩" are parts of the longer term "六合彩" (Mark Six Lottery). To address this problem, we try to achieve better boundary identification to filter out invalid terms by calculating the Asso-

[1] The Usenet newsgroups corpus is not used here because the bigram and *n*-gram information is not available.

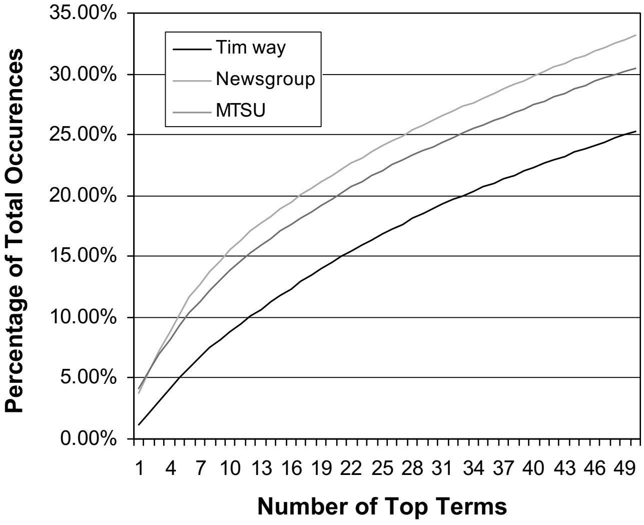

Fig. 2a. Comparison of the top 50 unigrams: the Timway search log data, the Usenet newsgroup corpus, and the MTSU corpus.

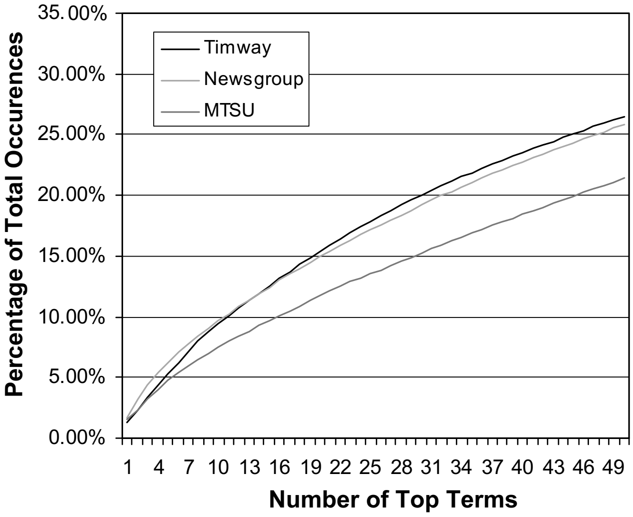

Fig. 2b. Comparison of the top 50 unigrams with function words removed: the Timway search log data, the Usenet newsgroup corpus, and the MTSU corpus.

ciation Norm Estimation (AE), Left Context Dependency (LCD), and Right Context Dependency (RCD) metrics for all the terms in our data (Chien, 1999). The Association Norm Estimation is calculated as follows:

$$\mathrm{AE} = \frac{f_x}{f_y + f_z - f_x} \tag{3}$$

where f_s is the frequency of a pattern s, x is a pattern to be estimated with length n, and y and z are the two longest composed substrings of x with length $n - 1$.

We calculate the Left Context Dependency and Right Context Dependency as follows:

$$\mathrm{LCD} = \frac{\mathrm{MAX}_{\alpha} f(\alpha x)}{f(x)} \tag{4}$$

$$\mathrm{RCD} = \frac{\mathrm{MAX}_{\beta} f(x\beta)}{f(x)} \tag{5}$$

where $\alpha \in L$ and $\beta \in R$, L and R being the sets of unique left and right adjacent character strings of x, respectively. Readers are referred to Chien (1999) for more details about these metrics.

Table 2a
The top 25 bigrams of Timway queries and MTSU corpus

Timway				MTSU (Da, 2004)			
Bigrams	Percentage (%)	Bigrams	Percentage (%)	Bigrams	Percentage (%)	Bigrams	Percentage (%)
香港	1.22	六合	0.24	一個	1.07	現在	0.30
成人	0.98	小說	0.24	什麼	0.86	出來	0.27
色情	0.70	漫畫	0.23	沒有	0.76	不能	0.25
公司	0.36	免費	0.23	自己	0.66	還是	0.25
下載	0.34	合彩	0.23	我們	0.62	不知	0.24
貼圖	0.29	手機	0.22	他們	0.60	可以	0.23
日本	0.28	明星	0.22	知道	0.42	女人	0.23
走光	0.28	電腦	0.22	起來	0.40	覺得	0.23
中國	0.27	二手	0.22	這個	0.40	因爲	0.22
電影	0.27	學生	0.20	時候	0.37	你們	0.22
情色	0.25	內衣	0.20	這樣	0.36	孩子	0.21
酒店	0.24	地圖	0.19	怎麼	0.32	那個	0.20
遊戲	0.24			已經	0.31		

Table 2b
The top 25 bigrams of Timway queries and MTSU corpus (function terms and incomplete terms removed)

Timway				MTSU (Da, 2004)			
Bigrams	Percentage (%)	Bigrams	Percentage (%)	Bigrams	Percentage (%)	Bigrams	Percentage (%)
香港	2.21	小說	0.43	知道	0.53	男人	0.16
成人	1.77	漫畫	0.42	女人	0.29	工作	0.16
色情	1.28	免費	0.41	孩子	0.26	聲音	0.15
公司	0.65	手機	0.40	眼睛	0.22	今天	0.15
下載	0.63	明星	0.40	心裏	0.22	時間	0.15
貼圖	0.53	電腦	0.40	東西	0.20	中國	0.15
日本	0.52	二手	0.39	先生	0.20	事情	0.14
走光	0.51	學生	0.37	看見	0.19	兒子	0.14
中國	0.50	內衣	0.36	父親	0.19	問題	0.14
電影	0.49	地圖	0.34	地方	0.19	太太	0.14
情色	0.46	中文	0.32	回來	0.19	告訴	0.14
酒店	0.44	鈴聲	0.31	生活	0.19	電話	0.13
遊戲	0.44			母親	0.17		

Based on these metrics, we filtered out the incomplete terms from the Timway data and obtained a set of valid terms.[2] Because we do not have the source documents in the MTSU data, we use the mutual information filter provided on their Web site to filter out the incomplete patterns.

[2] Of course, one should note that the filtering process is not perfect.

After removing the incomplete and invalid terms, we further removed the function words from both the Timway and MTSU data sets. The resulting data sets contain complete content words only. The top 25 bigrams obtained from these data sets are shown in Table 2b. As can be seen, there is only one overlapping bigram, being "中國" (China), in the top-25 lists. Among the top 100 most frequently occurring bigrams of the two corpora, there are only eight bigrams in common – "中國" (China) coming at the 9th place of the Timway bigram list and the 19th of the MTSU list, "電話"' (telephone) at the 26th of Timway and the 25th of MTSU, "學生" (student) at the 21st of Timway and the 36th of MTSU, "世界" (world) at the 43rd of Timway and the 38th of MTSU, "公司" (corporation) at the 4th of Timway and the 47th of MTSU, "日本" (Japan) at the 7th of Timway and the 78th of MTSU, "大學" (university) at the 57th of Timway and the 87th of MTSU, and "汽車" (automobile) appearing at the 63rd of Timway and the 99th of MTSU. The limited overlap of these two corpora further indicated that the character usage of query terms submitted by search engine users has features different from the character usage of common Chinese texts online, in terms of topics, patterns, and structures. The language of search engine queries has its unique characteristics, especially for Chinese search engines.

We also put the bigram frequencies in rank order and plotted their cumulative occurrences curves for comparison. Fig. 3a compares the cumulative occurrences of the top 5000 bigrams of Timway and those of MTSU without any preprocessing. This figure shows the patterns of the characteristics of the bigrams in their raw form. Fig. 3b shows the same curves after function terms and incomplete terms are removed.

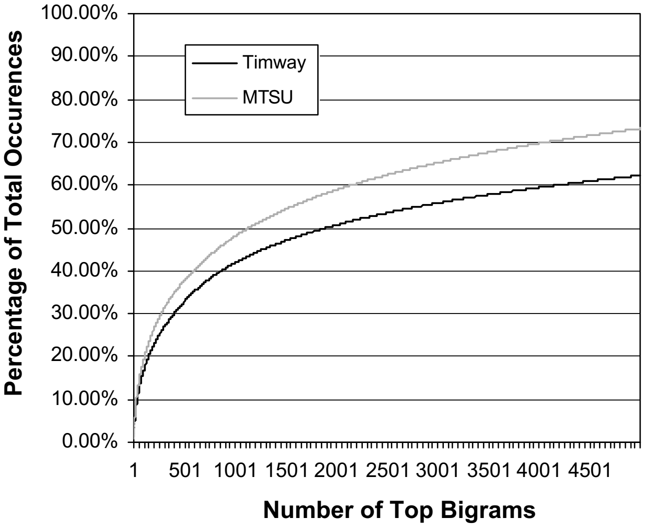

Fig. 3a. Comparison of the cumulated frequency curves for the top 5000 bigrams with the Timway search log data and the MTSU corpus (raw data).

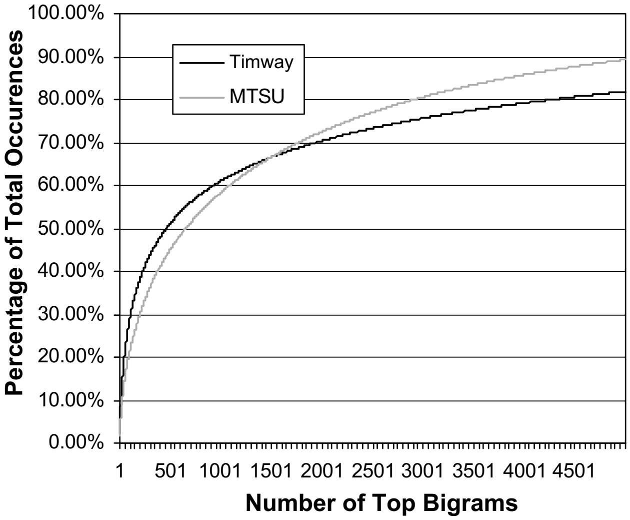

Fig. 3b. Comparison of the cumulated frequency curves for the top 5000 bigrams with the Timway search log data and the MTSU corpus (with incomplete terms and function terms removed).

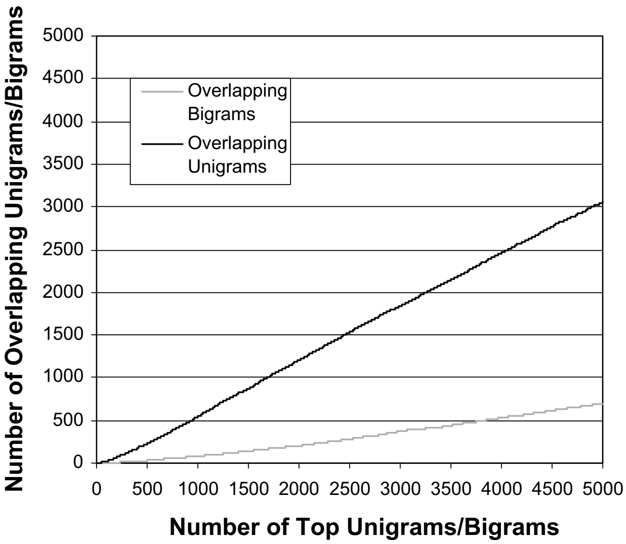

Fig. 4a. Overlapping unigrams and bigrams between Timway and MTSU (raw data).

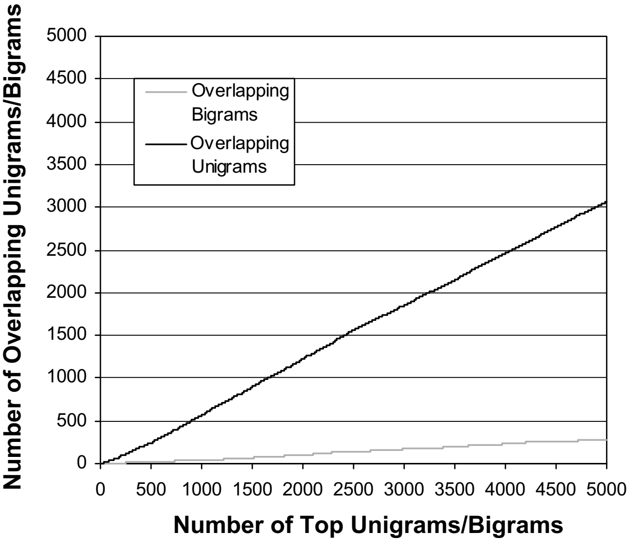

Fig. 4b. Overlapping unigrams and bigrams between Timway and MTSU (with incomplete terms and function terms removed).

Fig. 3a shows that when all bigrams are included, the MTSU data is less diversified than the Timway data. In Fig. 3b, where function terms and incomplete terms are removed, we can see that the Timway data is less diversified for the first 1500 bigrams, but become more diversified afterwards.

To further test our observation, we identified the overlapping unigrams and bigrams between the Timway corpus and the MTSU corpus in the range of the top 5000 from each corpus. Fig. 4 presents the result.

In our raw data, we find that these two corpora have 3056 overlapping unigrams (61.12%) out of the top 5000 unigrams (see Fig. 4a). The curve for overlapping bigrams increases much more gently than the one for overlapping unigrams. Only 698 bigrams (13.96%) out of the top 5000 bigrams from Timway appear in the top 5000 bigrams from MTSU. The great discrepancy between these two curves implies that although the two corpora have quite a lot of overlapping single characters, they have few common bigrams, which are a better representation for topics.

After removing incomplete terms and function terms, we obtain Fig. 4b. The number of overlapping unigrams increase slightly to 3063 out of 5000 (61.26%). For bigrams, the overlapping rate decreased to 5.62% (281 out of 5000). We suggest that the decreased overlapping of unigrams and bigrams is due to the fact that many common Chinese unigrams and bigrams are function words so the removal further reduced the overlap.

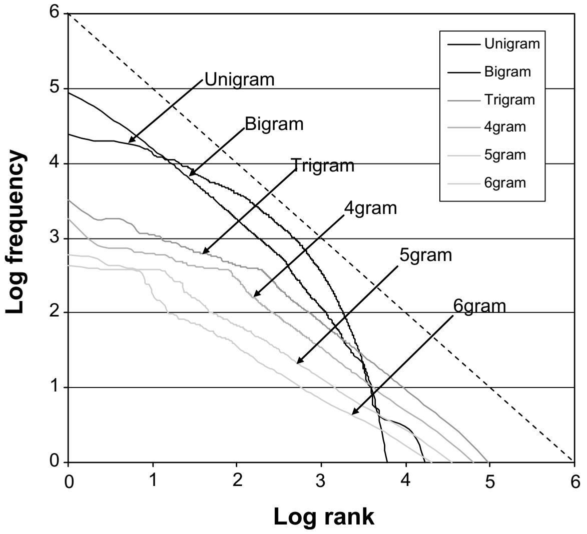

Fig. 5a. Zipf curves for the Timway corpus.

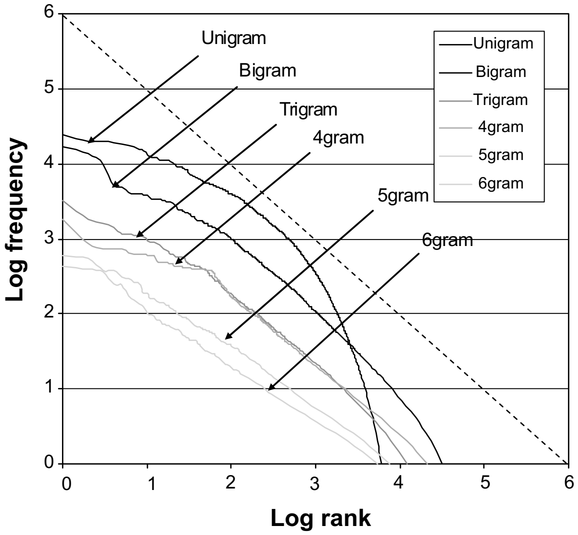

Fig. 5b. Zipf curves for the Timway corpus (with incompletle terms and function terms removed).

Table 3a
The value of β for best-fit straight line approximations to the Zipf curves

	Unigrams	Bigrams	Trigrams	4-grams	5-grams	6-grams
Timway	1.1416	1.1847	0.7100	0.7197	0.6779	0.6639

Table 3b
The value of β for best-fit straight line approximations to the Zipf curves (with incomplete terms and function terms removed)

	Unigrams	Bigrams	Trigrams	4-grams	5-grams	6-grams
Timway	1.1307	0.9163	0.8606	0.7941	0.7429	0.7482

5.3. Zipf distribution of Chinese search queries

Zipf distribution has been widely applied to analyze the distribution of terms used in Web search engines (Jansen et al., 2000; Spink et al., 2001; Wang et al., 2003). To test whether our data follow the same pattern, double-log rank-frequency plots were used to determine the accordance with a Zipf distribution. To plot the curve, the terms of interest is first ranked by their frequencies. The natural logarithm of a term's rank is then plotted against the natural logarithm of the term's frequency.

Since Chinese words consist of sequences of characters, it is interesting to analyze the characteristics of *n*-grams in search queries. Because there is no clear word boundary in Chinese (like the "space" counterpart in English), it is not easy to automatically extract the compound words (semantically complete words) in a Chinese query. Similar to our earlier analyses, we prepared two sets of data. In the first sets, we directly tokenized the search queries into characters (syllables) and all occurrences of any consecutive characters were extracted and the frequencies were recorded. It can be argued that most of the *n*-grams generated by this means are meaningless or semantically incomplete. However, it has been suggested that this is the nature of *n*-grams (O'Boyle, Owens, & Smith, 1994; Ney, 1999). Moreover, researchers have been using *n*-grams, which include semantically incomplete *n*-grams, with great success in modeling language over the last 20 years. Ha, Sicilia-Garcia, Ming, and Smith (2003) have also demonstrated the validity of this view and suggested that "every *n*-gram taken from a natural language text produced by humans has meaning, though often incomplete" (Ha et al., 2003, p. 8). Another reason for taking this method lies on the focus of our research. We analyze the character usage of Chinese search queries by character-based processing methods.

On the other hand, we applied the same set of metrics used earlier in order to obtain the second set of data. We calculated the Association Norm Estimation, Left Context Dependency, and Right Context Dependency for all the patterns in our queries and filtered our terms with a low value in these metrics.

For both sets of data, we calculated the frequencies of all *n*-grams ($n = 2$–6) and put them in rank order as we had done for the unigrams. The Zipf curves for *n*-grams for the Timway corpus are shown in Figs. 5a and 5b. In each Figure, the curves for different values of *n* are drawn and shown in the same chart in order to facilitate comparisons.

As illustrated in Figs. 5a and 5b, the *n*-gram Zipf curves can be represented by a single Mandelbrot form (Ha et al., 2002):

$$f = k/\gamma^{\beta} \tag{6}$$

where β is the magnitude of the negative slope of each line. The smaller the value of β, the more gently the line falls, indicating the relatively even usage of words in the corpus the curve represents. We show the values of β for the Timway data in Table 5 and the values of β after removing stop words in Table 5b.

We see that the values of β for unigrams (both Tables 3a and 3b) and bigrams (Table 3a) are higher than the standard magnitude of the negative slope of Zipf distribution ($\beta = 1$), showing that the relative frequencies of unigrams and bigrams are slightly higher than that of a standard Zipf distribution for high-ranking terms. The relatively low β-values of *n*-grams ($n = 3, 4, 5, 6$) indicate that the relative frequencies of words in *n*-grams are much lower than that of a standard Zipf distribution for high-ranking terms. In the following sub-sessions, we will discuss the characteristics of *n*-grams distribution in detail.

5.3.1. *Zipf distribution of unigrams*

In both Figs. 5a and 5b, the Zipf curve for unigrams is the most curved one of all, falling off very gently at the beginning and showing a rapidly sloping tail toward the end. It suggests that a small number of single characters occurred frequently and a large number of single characters occurred infrequently. This drop corresponds to the fact that the number of characters for low frequency increases comparatively slowly as the frequency decreases. It concurs with the findings reported in both the studies of Chinese search terms (Wang et al., 2003) and the studies of English search terms (Spink et al., 2001; Jansen et al., 2000). It implies that a more sophisticated model is needed to describe the query term distribution more accurately.

Table 4a
The top 10 *n*-grams in the Timway corpus

Trigrams		4-grams		5-grams		6-grams	
Freq	Term	Freq	Term	Freq	Term	Freq	Term
3177	六合彩	1812	有限公司	594	香港賽馬會	421	香港中文大學
1816	有限公	864	手提電話	541	倚天屠龍記	392	免費港股報價
1815	限公司	746	數碼相機	426	香港中文大	379	格蘭超級足球
1814	學生妹	736	情色文學	421	港中文大學	379	超級足球聯賽
1631	旅行社	726	成人漫畫	392	免費港股報	379	英格蘭超級足
1418	討論區	655	香港賽馬	392	費港股報價	379	蘭超級足球聯
1158	中原地	654	手機鈴聲	386	香港聯交所	371	網頁製作工具
1141	情色文	633	中原地圖	380	非牟利機構	256	城市電腦售票
1118	賽馬會	630	心理測驗	380	英格蘭超級	246	女子十二樂坊
1098	圖書館	625	明星合成	379	超級足球聯	173	香港名勝古蹟

Table 4b
The top 10 *n*-grams in the Chinese Timway corpus (with incomplete terms removed)

Trigrams		4-grams		5-grams		6-grams	
Freq	Token	Freq	Token	Freq	Token	Freq	Token
3177	六合彩	1812	有限公司	594	香港賽馬會	421	香港中文大學
1814	學生妹	864	手提電話	541	倚天屠龍記	392	免費港股報價
1631	旅行社	746	數碼相機	386	香港聯交所	371	網頁製作工具
1418	討論區	736	情色文學	380	非牟利機構	256	城市電腦售票
1158	賽馬會	726	成人漫畫	304	中國旅行社	246	女子十二樂坊
1098	圖書館	654	手機鈴聲	292	聖鬥士星矢	173	香港名勝古蹟
1049	貼圖區	633	中原地圖	263	香港學生妹	156	會議展覽中心
923	百老匯	630	心理測驗	256	公共圖書館	135	新世界傳動網
899	勞工處	625	明星合成	188	寵物小精靈	125	免費成人電影
896	模擬器	597	網頁素材	187	台灣成人網	103	香港有限公司

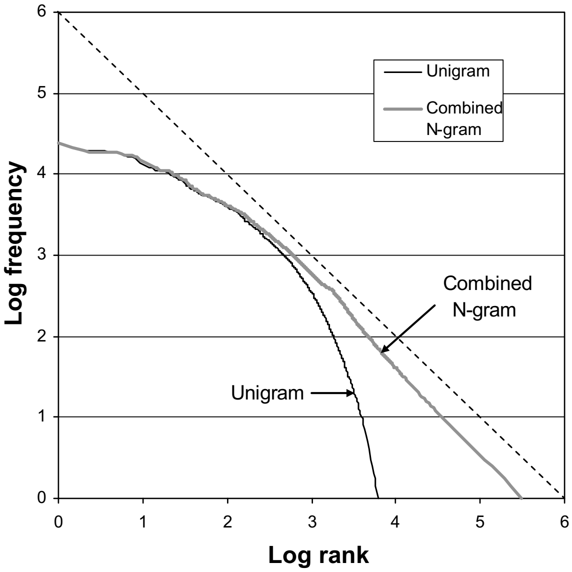

Fig. 6a. The unigram and combined *n*-gram Zipf curves for the Timway corpus.

5.3.2. *Zipf distribution of bigrams*

The data shown in Figs. 5a, 5b and Tables 3a, 3b suggest that the bigram curve follows the Zipf distribution best, except for the low ranking terms. The good accordance with Zipf distribution with the slope close to the theoretical value of 1 indicates that compound words consisting of two characters are frequently used in Chinese search query terms. The great discrepancy at the bottom of the curve suggests that there are a considerable number of infrequently occurring bigrams in the Chinese Web search engine.

5.3.3. *Zipf distribution of n-grams*

In Figs. 5a and 5b, the curves of trigrams, 4-grams, 5-grams and 6-grams are nearly parallel, showing similar structures and all four of them deviate from Zipf's law with smaller γ. Their separation is due to their different sizes. We can see from the Figures that the frequencies for the highest ranking terms always deviated from the straight lines and towards the horizontal lines. The ten most frequent *n*-grams in the Timway corpus are listed in Table 4a.

From Table 4a, we found that some of the top ten 5-grams and 6-grams are queries that refer to the same topic. Examples include the two incomplete terms "香港中文大" and "港中文大學", which are part of the term "香港中文大學" (The Chinese University of Hong Kong). Similarly, the incomplete terms "免費港股報" and "費港股報價" are parts

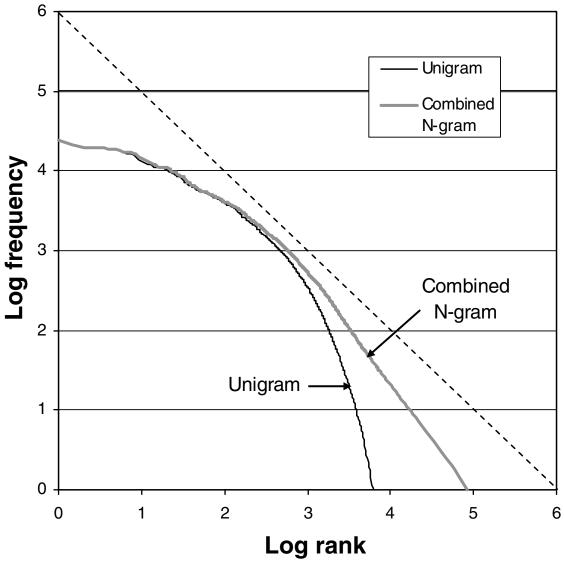

Fig. 6b. The unigram and combined *n*-gram Zipf curves for the Timway corpus (with incomplete terms and function terms removed).

of the term "免費港股報價" (free stock quote for Hong Kong stocks). To better understand the topics represented by the top *n*-grams, we remove the incomplete terms from the data using the AE, LCD, and RCD metrics discussed previously. The results are shown in Table 4b. From the Table, we can see some of the most frequently requested topics that are composed of 3 or more characters. These topics share some similarities to the top topics identified in the bigram analysis.

5.3.4. *Zipf distribution of combined n-grams*

As shown in Fig. 5, none of the *n*-gram curves follow Zipf's law perfectly. On the other hand, Ha et al. (Ha et al., 2003), discover that when all *n*-grams are combined together, including unigrams and semantically incomplete *n*-grams, Zipf's law is found to be approximately correct with a β-value close to 1 for all ranks. It led us to combine all *n*-grams and do the ranking and Zipf's distribution analysis to see whether our data has the same characteristics. We combined all the *n*-grams (with $n = 1–6$) into a single set and put them in rank order as we had done previously. Fig. 6 shows the resulting curve of combined *n*-grams with the unigram curve.

From Figs. 6a and 6b, we notice that the curves of combined *n*-grams and unigrams overlap at the beginning and fall off gently in the range of rank below 100. For rank greater than 100, the unigrams curve drops away from Zipf's slope of -1 rapidly, while the combined *n*-grams curve goes almost straight with slope close to -1 (especially in Fig. 6a). The deviation of the unigrams was made up by the other *n*-grams. Different from the results of Ha et al.'s research on the extension of Zipf's law to general Chinese text, the combined Zipf curve for our data follows Zipf's slope only when rank is greater than about 100. The phenomenon that a small number of single characters occurred frequently in query logs may account for the discrepancy.

6. Discussions

This article is among the few studies investigating the characteristics of Web searching in Chinese. Some valuable findings on non-English information Web search behavior are revealed in this article. By comparing the character usage of our corpus with that of two corpora obtained from general online texts, we found that our data had the lowest growing rate for the distribution of occurrences. We speculate that people tend to use more and more diverse set of terms when searching online than writing online texts. The little overlapping in high-frequency characters between our corpus and that of the two corpora further confirmed that the usage of characters for search queries has distinctive features in several aspects. The language of queries has its unique characteristics and it should be further investigated in order to improve the design and development of search engines.

We also found that the bigram search terms in a Chinese Web search engine follow the Zipf distribution very closely. They are more topic related and can indicate what types of information are needed. The popular use of a small number of bigrams suggests that information content providers can expect to reach Web users by targeting specific high frequency terms to make these resources easy to access.

Our study utilized Zipf distribution to undertake *n*-gram analysis and noticed that the curves of trigrams, 4-grams, 5-grams and 6-grams had similar structures with β-values of about 0.7 and were greatly unbalanced for the high ranking terms. The findings concur with previous studies which have been reported by Ha et al. (2002). We also listed the 10-highest frequency *n*-grams ($n = 3–6$) and further figured out some information about what users were interested in online. In addition, we combined all the *n*-grams and found that the distribution approximately obeys Zipf's law for ranks above 100.

Generally, the language of Chinese Web queries has its own unique characteristics and more research is needed. However, there are several limitations of the current study that should be noted. First, the lack of previous studies on search log analysis in Chinese and the lack of available data on Chinese search query make it impractical for us to compare our data with analogous corpora. Also, the fact Chinese are spoken and written differently in different areas (e.g., Traditional Chinese characters in Hong Kong and Taiwan versus Simplified Chinese in mainland China) may have impacts on the character usage. Our study only focused on a search engine in Hong Kong and further research is needed to study such impacts. In addition, since our data only cover three months, the short period might weaken the generalizability of our findings. It may affect the lists of top unigrams and *n*-grams due to the seasonal patterns and topic shift over time (Chau et al., 2005). A longitudinal study on Chinese search queries concerning seasonal impacts will be an interesting area for future research.

Another limitation concerns the comparison of search queries with other online Chinese corpora due to their distinct features. A search engine accepts queries with few restrictions while newsgroups and the general online sources are often maintained and monitored by humans (who may reject postings on particular topics). This should be noted when interpreting the results of the current study.

7. Conclusion and future directions

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

As discussed earlier, our study also found that the resulting distributions of *n*-grams ($n = 1–6$) deviated from the standard Zipf distribution for the high and low ranking terms, indicating that we may need a more sophisticated model to describe the distribution of Web search queries, especially for non-English search engines.

Another direction of future research is to study the character usage in Chinese search queries in different areas such as mainland China and Taiwan. Because of the different vocabularies and the different dialects spoken, their characteristics could be different. Also, the current research only looked at Chinese search queries. It would be interesting to analyze the Web search queries in other Asian languages.

It is also important to note that as the usage of "mixed words" (words that consist of both Chinese and English) is getting more popular in daily life; the amount of mixed queries submitted by users has also increased. It would be valuable to investigate the characteristics of mixed queries and their implications in Web searching and search engine design. In addition, the study on the correlation and patterns between the queries in the two languages submitted to Timway will be interesting and meaningful.

Acknowledgements

This research has been supported in part by a Seed Funding for Basic Research (PI: M. Chau) from the University of Hong Kong. We would like to thank Timmy Yu from Timway Hong Kong Search Engine Limited for his help in providing the search log data used in this study. We also thank Jackey Ng, Raygen Lam, and Ivan Ho from the University of Hong Kong for their help in data processing.

References

Chau, M., Fang, X., & Liu Sheng, O. R. (2005). Analysis of the query logs of a web site search engine. *Journal of the American Society for Information Science and Technology, 56*(13), 1363–1376.

Chau, M., Fang, X., & Yang, C. (2007). Web searching in Chinese: A study of a search engine in Hong Kong. *Journal of the American Society for Information Science and Technology, 58*(7), 1044–1054.

Chau, M., Qin, J., Zhou, Y., Tseng, C., & Chen, H. (2005). 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*.

Chien, L.-F. (1999). PAT-tree-based adaptive keyphrase extraction for intelligent chinese information retrieval. *Information Processing and Management, 35*, 501–521.

Croft, W. B., Cook, R., & Wilder, D. (1995). Providing government information on the internet: Experiences with THOMAS. In *Proceedings of the digital Libraries'95 conference, Austin, Texas* (pp. 19–24).

Da, J. (2004). Chinese text computing. <http://lingua.mtsu.edu/> Accessed 05.11.05.

Fedorowicz, J. (1982). A Zipfian model of an automatic bibliographic system: An application to MEDLINE. *Journal of American Society of Information Science,* (33), 223–232.

Ha, L. Q., Sicilia-Garcia, E. I., Ming, J., & Smith, F. J. (2002). Extension of Zipf's law to words and phrases. In *Proceedings of the 19th international conference on computational linguistics* (pp. 315–320).

Ha, L. Q., Sicilia-Garcia, E. I., Ming, J., & Smith, F. J. (2003). Extension of Zipf's law to word and character *n*-grams for English and Chinese. *Computational linguistics and Chinese Language Processing, 8*(1), 77–102.

Hölscher, C. (1998). 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*.

Huang, C. K., Chien, L. F., & Oyang, Y. J. (2003). 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), 638–649.

Huang, C. K., Oyang, Y. J., & Chien, L. F. (2001). A contextual term suggestion mechanism for interactive search. In *Proceedings of the first web intelligence conference (WI'2001), Japan* (pp. 272–281).

Jansen, B. J., & Pooch, U. (2000). Web user studies: A review and framework for future work. *Journal of the American Society of Information Science and Technology, 52*(3), 235–246.

Jansen, B. J., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: A study of user queries on the web. *ACM SIGIR Forum, 32*(1), 5–17.

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 and Management*(36), 207–227.

Jones, S., Cunningham, S. J., & McNam, R. (1998). Usage analysis of a digital library. In *Proceedings of the 3rd ACM conference on digital libraries, Pittsburgh, PA, USA, June* (pp. 293–294).

Mandelbrot, B. (1953). An information theory of the statistical structure of language. In Jackson Willis (Ed.), *Communication theory* (pp. 486–502). New York: Academic Press.

Nelson, M. J. (1989). Stochastic models for the distribution of index terms. *Journal of Documentation, 45*(3), 227–237.

Ney, H. (1999). The use of the maximum likelihood criterion in language modelling. In K. Ponting (Ed.), *Computational models of speech pattern processing* (pp. 259–279). Berlin, Germany: Springer.

O'Boyle, P., Owens, M., & Smith, F. J. (1994). A weighted average *n*-gram model of natural language. *Computer Speech and Language*(8), 337–349.

Pu, H. T., Chuang, S.-L., & Yang, C. (2002). Subject categorization of query terms for exploring web users' search interests. *Journal of the American Society for Information Science and Technology, 53*(8), 617–630.

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

Smith, F. J., & Devine, K. (1985). Storing and retrieving word phrases. *Information Processing & Management, 21*(3), 215–224.

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From E-sex to E-sommerce: Web search changes. *IEEE Computer, 35*(3), 107–109.

Spink, A., & Ozmultu, H. C. (2002). Characteristics of question format web queries: An exploratory study. *Information Processing and Management*(38), 453–471.

Spink, A., Ozmutlu, H. C., & Lorence, D. P. (2004). Web searching for sexual information: An exploratory study. *Information Processing and Management*(40), 113–123.

Spink, A., Wolfram, D., Jansen, B. J., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology, 52*(3), 226–234.

Tsai, C.-H. (1996). Frequency and stroke counts of Chinese characters. %3chttp://technology.chtsai.org/charfreq/%3e Accessed 20.05.05.

Wang, P., Berry, M. W., & Yang, Y. (2003). Mining longitudinal web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology, 54*(8), 743–758.

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.

Wolfram, D., Spink, A., Jansen, B. J., & Saracevic, T. (2001). Vox Populi: The public searching of the web. *Journal of the American Society for Information Science and Technology, 52*(12), 1073–1074.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley.