Webology, Volume 2, Number 3, October, 2005

How Do Search Engines Handle Chinese Queries?

Haidar Moukdad
Assistant Professor, School of Information Management, Dalhousie University, Halifax, Nova Scotia B3H 3J5 Canada Phone: 1-902-494-2462

Hong Cui
Information Specialist, Access Copyright-The Canadian Copyright Licensing Agency, 1 Yonge Street, Suite 1900, Toronto, Ontario M5E 1E5 Canada Phone: 1-416-868-1620 ext. 339

Received September 13, 2005; Accepted October 16, 2005

Abstract

The use of languages other than English has been growing exponentially on the Web. However, the major search engines have been lagging behind in providing indexes and search features to handle these languages. This article explores the characteristics of the Chinese language and how queries in this language are handled by different search engines. Queries were entered in two major search engines (Google and AlltheWeb) and two search engines developed for Chinese (Sohu and Baidu). Criteria such as handling word segmentation, number of retrieved documents, and correct display and identification of Chinese characters were used to examine how the search engines handled the queries. The results showed that the performance of the two major search engines was not on a par with that of the search engines developed for Chinese.

Keywords

World Wide Web, Search engines, Chinese language

Introduction

Over the last decade, the World Wide Web (Web) has become a major source of information. Although English is still the dominant language on the Web, information in other languages is steadily gaining prominence. Among non-English languages, the use of Chinese has significantly grown in the past few years. In 1995, China opened two 64K specialized lines in Beijing and Shanghai and made the Internet accessible (CNNIC, 2003). In 2002, Chinese became the second most often used online language following English (Cyber Atlas, 2003). According to recent news released by China Internet Network Information Center (CNNIC), the number of Internet users in China reached 103 million at the end of June 2004, 35 million more than half a year earlier. The number of Internet users doubles nearly every six months and the number of websites every year. Considering the large population of China [1,304,196 in 2003] (UNICEF, 2005) and the rocketing sale trends of PCs and mobile phones, the number of Internet users and the level of need for information in Chinese will keep rising.

Despite the increasing availability of information in Chinese on the Web, there have been only a few attempts at exploring issues related to how search engines handle this language and to the effects of the characteristics of Chinese on information retrieval (IR). This paper outlines the characteristics of the Chinese language that affects IR and explores how major search engines handle queries entered in Chinese.

Characteristics of Chinese

Modern written Chinese originated from the inscriptions on bones and tortoise shells 3,000 years ago (Shan Dynasty, 16th century to 11th century B. C.). The initial characters were imitations of various objects. With the development of the language, words reflecting abstract concepts were created by combining or symbolizing the concrete concepts. For example, sun (day) was written as (日), and moon (month) as (月). To express the concept of bright or brightness, (明) was created by putting sun and moon together. In other cases, part of the character was used to mark the pronunciation or to indicate the meaning (Crystal, 1987).

Modern Chinese has simplified the pictographs by using characters made up of seven strokes (horizontal and vertical strokes, left-falling and right-falling strokes, a point stroke, and a hook stroke). No word boundary is necessary between two characters in Chinese. Therefore, a sentence is a continuous string of different characters, without spaces, that ends with a punctuation mark (CJK input methods, 2003).

Another significant characteristic of Chinese is that it does not have variations of words: no changes of tenses and no plural forms. The reader has to find clues of tense and quantity from the context and the quantitative words. Also, verbs are not always required in a sentence.

Though acronyms exist in Chinese, they are not necessarily extracted from the initials of each word, but from some characters of a phrase. For example, UNESCO is translated to 联合国教育科学文化组织 in Chinese, and the acronym is 联合国教科文组织. If there are some repeated characters in a phrase, the quantitative words are used to simplify the phrase. For example, 前生今生来生 is shortened to 三 (three) 生. Sometimes, the initial character is extracted from a word, such as 美 (美国, the United States), 加 (加拿大, Canada; 加纳, Ghana), and 日 (日本, Japan). However, these characters have to be carefully understood in their context because 美 can also mean beautiful in Chinese; 加 plus, add, put; and 日 sun and day.

Some words can be used as auxiliary words and notional words as well. One good example is " 的 ". When it is pronounced as "de", it means "of"; when it is pronounced as "di", it means the central target of archery; when it is pronounced as "di", it means taxi. In the last two cases, 的 is an independent notional word (Crystal, 1987).

Some linguists in the last century tried to abandon the Chinese characters and follow the Roman alphabet. In 1950s, Mainland China simplified some strokes of Chinese characters and formed Simplified Chinese; Pinyin has also been introduced to help the pronunciation and to replace the characters in the future. Although trying to use the Roman alphabet was a failure, Pinyin has been accepted as a useful tool to learn the pronunciation of characters. Meanwhile, Traditional Chinese has continued to be used in Hong Kong and Taiwan. For pronunciation, the Wade-Giles (WG) system has been used (Chinese Romanization Guide, 2003). For historical purposes, most North American bibliographic records had been in the WG system until the Library of Congress started its Pinyin Conversion Project in 2000 (Library of Congress, 2003). Nowadays, the co-existence of the two transliteration systems can be seen in many English sources related to China.

The simultaneous usage of Simplified and Traditional Chinese characters in the Mainland, Hong Kong and Taiwan brings some problems to the online world. Two character-set coding schemes, GB and Big 5, are used to encode Simplified and Traditional Chinese. These two coding schemes are not compatible, and therefore require the user to install both schemes to view them.

Some new words, especially concepts from foreign languages or technical terms, are written differently in Simplified and Traditional Chinese. One example is the word "Internet", which is translated to 互联网 or 因特网 in Simplified Chinese, and to 国际网路 in Traditional Chinese.

Previous work

On the multilingual Web, the characteristics of languages have to be considered in order to efficiently retrieve information. Grefenstette (1998) discussed language identification methods (n-grams and short word techniques), techniques for automatically generating queries in other languages (stemming and morphological analysis; dictionary-based translation, corpus-based translation, machine-translation based; query conflation), and retrieving and merging query results (weighting schemes).

Pirkola (2001) proposed a morphological classification of languages from an IR perspective, and focused on showing the differences among the morphologies of languages and their effect on IR in general and on cross-language IR (CLIR) in particular. Pirkola emphasized the differences in the frequency of derivatives and compound words and in the use of inflectional morphemes to create new words, and presented a morphological typology for IR. He also asserted the need for semantic and syntactic typologies, basing his assertion on the problems created by lexical ambiguities and by the rules of word orders.

Several studies have been conducted on the effectiveness of search engines in handling languages other than English. Bar-Ilan & Gutman (2003) tested general (English oriented) and local (non-English oriented) search engines on handling queries in Russian, French, Hungarian, and Hebrew. The authors ran queries in all four languages in the local and the general search engines, and found that in most cases the latter ignored the special characteristics of the language of the queries and did not properly handle diacritics. This resulted in a high number of missed documents; local engines would have retrieved these documents, because they would have handled the special characteristics of their respective languages and their diacritics. Bar-Ilan and Gutman concluded that morphological variations among languages must be considered by the developers of search engines, and users should be made aware of what they miss when they use the general search engine to find information in languages other than English.

Using a local version of AltaVista, Moukdad & Large (2001) created a test database of 271 Arabic HTML documents and ran 560 queries against this database. The queries were constructed to highlight the special characteristics of Arabic, especially the occurrence of prefixes, which are very common in Arabic words. The results of the searches showed that AltaVista search features and its indexing algorithms did not handle Arabic queries well, and did not provide any mechanisms to address Arabic prefixes. The need to develop new algorithms to handle Arabic prefixes was suggested as a way to foster more research on the topic and to identify potential problems for CLIR.

Foo & Li (2004) conducted experiments to study the impact of Chinese word segmentation and its effect on IR. Four automatic character based segmentation approaches and a manual one were used to index and evaluate the accuracy of these approaches. The experiments revealed that the segmentation approach had an effect on IR effectiveness. Better IR results could be achieved by using the same method for query and document processing, which increased the probability of matching queries to documents.

Method

A set of 10 queries were created and entered in two major search engines (Google and AlltheWeb) and two search engines developed for Chinese (Sohu and Baidu). The queries were based on terms that were selected from a Chinese-English dictionary to emphasize the characteristics of Chinese as explained above. The ten queries were: 吧, 巴, 的, 的哥, 和, 或, 加, 山大, 三生, and 电脑. Some of the terms used in these queries have different meanings, as they can belong to more than one grammatical category. The following table lists the 10 queries, the grammatical categories they belong to, and their English translations.

Table 1. The 10 queries used in this study

Query	Grammatical categories and English translation
吧	**conjunction**, (used at the end of a sentence) indicating entreaty; suggestion; command; etc.; **noun**, transliteration from bar; means bar; inn, etc.
巴	**noun**, a snake species; an ancient nation in south-east China; **verb**, to stick to
的	**noun**, taxi; central target of archery
的哥	**noun**, male taxi driver
和	**conjunction**, and; **adjective**, kind; **noun**, peace; total; **preposition**, to; **verb**, to follow; to mix; to stir
或	**conjunction**, or; **pronoun**, somebody
加	**verb**, to add; to put on; to impose; to take part in; to bully; **adverb**, more; **noun**, benefit; good; short for Canada and Ghana
山大	**noun**, short for Shandong University (山东大学) or Shanxi University (山西大学)
三生	**noun**, three students
电脑	**noun**, computer

As mentioned above, the selected search engines were the Chinese version of Google, AlltheWeb (Simplified Chinese was chosen as language preference), Sohu, and Baidu. As the latter two search engines are based in Mainland China and, therefore, use Simplified Chinese (For residents in Hong Kong and Taiwan, they provide local editions in Traditional Chinese), only Simplified Chinese searches were conducted in Google and AlltheWeb (Google and AlltheWeb provide choices to search in Simplified Chinese, Traditional, or both).

All search engines search for exact characters only. Google claims that it can combine Simplified and Traditional if the user chooses to search in Chinese. To indicate the segmentation of different keywords, "space" or "+" are required instead of "AND"; a quotation mark is not required when searching for phrases because there are no spaces between words in Chinese. Google and AlltheWeb use the Unicode scheme (for both Simplified and Traditional Chinese), while Baidu and Sohu use the GB scheme.

The 10 queries were entered in each of the four search engines, and the results were saved and examined to explore the performance of each engine based on three criteria: number of retrieved documents and search performance, word segmentation, and correct display of Chinese characters. As it was not possible to measure recall and precision in this search environment, the search performance of engines was assessed in relation to the number of retrieved documents, taking into account the different sizes of indexes and the possibility that Baidu and Sohu applied morphological analysis on queries to retrieve more documents.

Results

Number of retrieved documents and search performance

Table 2 shows the number of documents retrieved by the four engines for each query, and includes, where appropriate, comments on the searches in specific engines. For example, quotation marks were used in Google for query 的哥.

In general, Baidu performed well in each query. For single character queries, except 的, Sohu had the second largest number of hits, but it fell behind Google and AlltheWeb in some phrase searches. Google did not support the search of auxiliary words; therefore, 的 could not be searched, and when 的哥 was searched as a phrase, 的 was also filtered out of the phrase: the use of the quotation marks was not enforced. Sohu supported the search of auxiliary words, but 的 was not on its searching list.

Table 2. Queries and the number of retrieved documents

Query	Google	AlltheWeb	Baidu	Sohu
吧	3,910,000	7,786,923	25,100,000	5,432,327
巴	574,000	1,307,355	7,870,000	885,495
的	N/A	60,302,661	231,000,000	N/A
的哥	56,000 Quotation marks required	66,362 Exact form	107,000 Exact form	58,607 Exact form Provides spelling choice of "低格"
和	6,640,000	37,163,603	110,000,000	32,900,562

或	4,610,000	19,272,258	61,900,000	14,484,466
加	2,650,000	6.045,707	22,600,000	3,888,790
山大	43,100 Not exact form	43,794 Exact form	109,000 Exact form	43,104 Exact form
三生	31,500	45,602	163,000	28,877 Provides alternative spellings of "三盛" "三声"
电脑	4,470,000 Includes both电脑 and 计算机; with quotation marks, 4,280,000 hits	6,722,568 Exact form	37,500,000 Exact form	4,483,390 Exact form

AlltheWeb had problems with the correct identification of retrieved documents. As shown in Figure 1, although the search was limited to Chinese in its two forms (Simple and Traditional), the first document retrieved with the query of 巴 was in Korean; other documents retrieved using the same query were also in Korean.

Figure 1: Korean document retrieved by AlltheWeb

Word segmentation

Appropriate segmentation is still a problem for search engines, although Baidu and Sohu, the local engines, fared better than Google and AlltheWeb.

All four search engines had problems in identifying correct word segmentations. For the query 三生, "高三生" [a compound word: 高三: grade 3 of high school +生 student(s): grade 3 high school student(s)] appeared in the first ten hits of Google (One hit is shown in Figure 2) and AlltheWeb search, and in the 19th and 134th hits of Sohu and Baidu respectively, where "初三生" (grade 3 junior school students) and "大三生" (grade 3 college students) were chosen. Similarly, there were some problems in segmenting 山大. AlltheWeb and Baidu's 10th hit were 东阳山大酒店 and 亚力山大, which should be segmented as 东阳山+大酒店 (Mountain Dongyang + grand hotel) and 亚力山大 (the transliteration of Alexander). Google and Sohu encountered the same problems in their 11th and 13th hits: 盘龙山庄大酒店 (盘龙+山庄+大酒店=Panlong Cottage Grand Hotel) and 普陀山大酒店 (普陀山+大酒店=Mountain Putuo Grand Hotel). In this case, Google's 11th hit embedded another character "庄" (village, cottage) between the two characters 山 and 大 as shown in Figure 3.

Figure 2: Google's hit of 三生 was actually 高三生

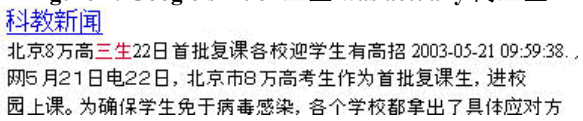


Figure 3: Google's error in segmenting 山大

Character display

For 电脑 (computer), Google considered 电脑 a Traditional Chinese word, while 计算机 was treated as a Simplified Chinese word, and included both words in its hits. In fact, the two words are used as synonyms. Both Google and AlltheWeb could not correctly display at least one hit from http://www.cfan.net.cn/ (Figures 4 and 5). Google had two hits of this URL: it displayed correctly the second hit, but had display problems similar to AlltheWeb's in the first hit.

Figure 4: Wrong character display in AlltheWeb when searching 电脑

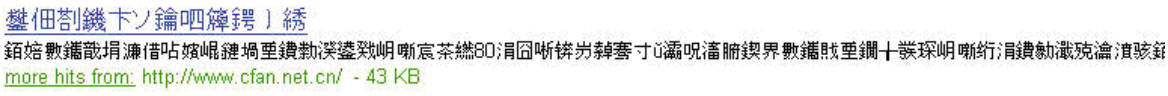
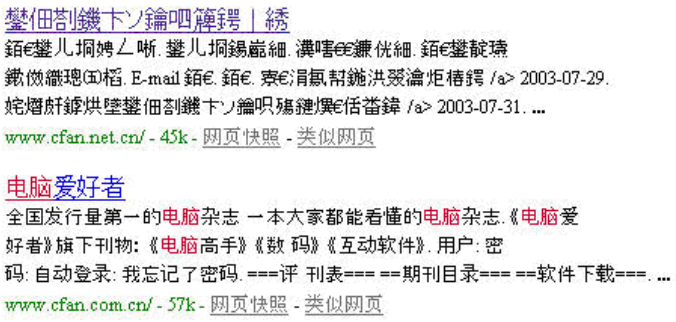


Figure 5: Wrong character display in Google when searching 电脑



Conclusion

As the Web continues to become more multilingual, and as languages other than English continue to gain ground on the Web, the need to develop search engines to handle all these languages has become more apparent. Documents published in languages that do not share the linguistic characteristics of English are more likely to be missed or improperly indexed by major search engines than English documents.

This exploratory study showed that major search engines designed for English do not handle Chinese queries as well as search engines specifically designed for Chinese do. Research on other languages has reached similar results, pointing to the need for new approaches to IR issues on the Web and for serious investigations of the feasibility of developing super search engines capable of handling a multitude of languages with equal degrees of effectiveness and efficiency. For the Web to be truly multilingual and truly accessible to people of different linguistic backgrounds, its main information locators, the major search engines, must evolve and go beyond English to accommodate other languages.

It might be unrealistic at this point to suggest that search engines should be equipped with all the linguistic tools capable of handling all languages, but a gradual progress towards this goal is not far fetched. The developers of search engines should seriously start looking at implementing the basic requirements of truly multilingual engines, starting with accommodating the morphological differences among languages and moving towards handling syntax rules and word segmentation mechanisms. There is a growing body of research on the morphologies of different languages and their effect on IR; the developers of search engines should pay attention to the results of this research and use them towards the development of engines with capabilities to handle queries with the same degree of effectiveness regardless of the language used.

References

- Bar-Ilan, J. & Gutman, T. How do search engines handle non-English queries? A case study. Retrieved July 28, 2003, from http://www2003.org/cdrom/papers/alternate/P415/BARILAN.HTM

- China Internet Network Information Center (CNNIC). Retrieved December 12, 2004, from http://www.cnnic.net.cn/en/index/0O/index.htm
- China Internet Network Information Center (CNNIC). (2003). The Internet Timeline of China Part I. Retrieved January 10, 2004, from http://www.cnnic.net.cn/html/Dir/2003/12/12/2000.htm
- Chinese Romanization Guide. Retrieved July 28, 2003, from http://www.edepot.com/taoroman.html
- CJK input methods. Retrieved July 28, 2003, from http://www.geocities.com/fontboard/cjk/input.html
- Crystal, D. (1987). Chinese culture studies: the Chinese language and writing. *The Cambridge Encyclopedia of Language*. Retrieved July 28, 2003, from http://acc6.its.brooklyn.cuny.edu/~phalsall/texts/chinlng2.html
- Cyber Atlas. Web pages by language. Retrieved July 27, 2003, from http://cyberatlas.internet.com/big_picture/demographics/article/0,,5901_1487151,00.html#table2
- English, French, German, and Chinese Romanisations of Chinese. Retrieved July 28, 2003, from http://www.sinistra.net/els/sup/transcript.html
- Foo, S. & Li, H. (2004). Chinese word segmentation and its effect on information retrieval. *Information Processing & Management*, 40 (1), 161-190.
- Grefenstette, G. (1998). Problems and approaches to cross language information retrieval. *Proceedings of the ASIS Annual Meeting*, 35, 143-152.
- Library of Congress. Pinyin Conversion Project. Retrieved July 27, 2003, from http://www.loc.gov/catdir/pinyin/pinyin.html
- Moukdad, H. & Large, A. (2001). Information retrieval from full-text Arabic databases: Can search engines designed for English do the job? *LIBRI*, 51 (2): 63-74.
- Pirkola, A. (2001). Morphological typology of languages for IR. *Journal of Documentation*, 57 (3), 330-348.
- UNICEF. At a glance: China. Retrieved February 21, 2005, from http://www.unicef.org/infobycountry/china.html

---

Bibliographic information of this paper for citing:

Moukdad, H. & Cui, H. (2005). "How Do Search Engines Handle Chinese Queries?" *Webology*, **2** (3), Article 17.
Available at: http://www.webology.org/2005/v2n3/a17.html

This article has been cited by other articles.
