Analysis of the Bilingual Queries of a Chinese Web Search Engine

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

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

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

Christopher C. Yang
Dept. of SEEM
The Chinese Univ. of HK
Shatin, Hong Kong
yang@se.cuhk.edu.hk

Abstract

In this research, we report our analysis of the bilingual (Chinese and English) search queries collected from a Chinese Web search engine. In particular, we classified the mixed queries into a few categories and tried to explore users' behaviors when conducting searches.

Keywords: Web Mining, Bilingual Search Queries, Search Behavior

1. Introduction

In cultures where people use both Chinese and English in their daily lives, the usage of mixed terms is prevailing. It is reflected by the phenomenon that people input mixed query terms to search in Web engines that support multiple languages. However, little is known about why and how people use mixed queries. We address these issues in this research.

2. Related Studies

The increasing growth of online population of non-English speakers has drawn great research interests in analyzing Web users' behavior in various languages (Wang et al., 2003; Rieh and Rieh, 2005; Petrelli et al., 2004). Petrelli et al. (2004) found that search behaviors were dependent upon user goals and purposes for searching as well as on "language knowledge of individuals and cognitive demands of the cross-language task itself" (Petrelli et al., 2004, p. 928). Rieh and Rieh (2005) verified this conclusion. Lewandowski (2006) classified about 1,500 queries from three German search engines and compared their results with those of Spink et al. (2002).

3. Data and Method

The query log used in our study covers a three-month period from December 1, 2003 to March 2, 2004 for a search engine in Hong Kong called Timway. The data consist of 1,255,633 queries in total. Out of these queries, 536,814 are Chinese queries, 641,169 are English queries, and 77,650 are mixed queries (Chau et al., forthcoming). The data were preprocessed and the bilingual queries were extracted for analysis in this study.

4. Analysis Findings

Based on our analysis, the Chinese-English mixed search queries can be classified into six types. The first type includes names of magazines, places, and firms such as "東touch" (East Touch), "UA時代廣場" (UA Times Square), and "ACM又一城" (ACM Festival Walk). The second type includes mixed queries that are used because the English part of the queries does not have a popular Chinese translation and thus cannot be substituted by simple Chinese translation. Examples are “mp3”, “bt”, “dvd”, “midi”, and “ICQ”. Most of them are related to computer technologies. For the third type, although the English part of the query has proper translation in Chinese, it was still entered as English by the user. A possible cause is the culture in Hong Kong, where people are used to adding English words into both Chinese writing and speaking. They habitually combine Chinese and English together to form a complete search query. For instance, “成人game” (adult game) and “明星wallpaper” (wallpaper of celebrities) etc. In the fourth type of mixed queries, the English parts are the abbreviations of certain English phrases and are popularly used by people. For example, “3G” in“3G手機” (3G mobile phone) is the shortened form of “3rd Generation”; “IQ” in “IQ題” (IQ questions) refers to “Intelligence quotient”; and “AV” in “日本AV” (Japan AV) stands for “adult video”. The fifth type of queries include those consisting of Chinese words and their corresponding English terms which have the same meaning, such as “Yuen Long 元朗” “Bowie Lam 林保怡”. The users might intent to get a high recall rate in these cases. The sixth type of mixed queries are those made up of an English brand name and a Chinese product name, such as “Canon鏡頭” (Canon lens) “Sharp手機” (Sharp cell phone), and “Panasonic數碼攝錄機” (Panasonic digital video camera). The top 10 mixed queries are shown in Table 1.

Rank	Bilingual Search Queries	Frequency	Rank	Bilingual Search Queries	Frequency
1	頭文字D	98	6	AV女優	64
2	BT下載	96	7	H漫	62
3	mp3機	91	8	h漫	54
4	bt下載	86	9	H漫畫	53
5	卡拉ok	84	10	外星BB撞地球	53

Table 1: Top 10 Mixed Queries

5. Future Work

Further analysis of the bilingual queries will be conducted. Currently, we are planning to sample a subset of queries and ask domain experts to manually classify them into these categories. Moreover, to speedup the searching process of mixed terms, we could extract possible key terms from the set of frequently used queries and try to obtain expansion of synonymous terms in advance. Our findings could help in the design of Web search engines.

References

1. Chau, M., Fang, X., and Yang, C. C. “Web Searching in Chinese: A Study of a Search Engine in Hong Kong,” *Journal of the American Society for Information Science and Technology*, accepted for publication, forthcoming.
2. Lewandowski, D. “Query Types and Search Topics of German Web Search Engines Users,” *Information Services and Use* (26), 2006.
3. Petrelli, D., Beaulieu, M., Sanderson, M., Demetriou, G., Herring, P., & Hansen, P. “Observing users, designing clarity: A case study on the user-centered design of a cross-language information retrieval system,” *Journal of the American Society for Information Science and Technology* (55), 2004, 923–934.
4. Rieh, H. Y., and Rieh, S. Y., “Web Search across Languages: Preference and Behavior of Bilingual Academic Users in Korea,” *Library & Information Science Research* (27), 2005, pp. 249-263.
5. Spink, A., Ozmutlu, S., Ozmutlu, J.C., & Jansen, B.J. “U.S. versus European Web Searching Trends,” *SIGIR Forum*, (36:2), 2002, pp. 32-38.
6. 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.