# Determining the User Intent of Chinese-English Mixed Language Queries Based On Search Logs

Hengyi Fu, Florida State University
Shuheng Wu, City University of New York

**Abstract**
With the increasing number of multilingual web pages on the Internet, multilingual information retrieval has become an important research topic. While queries are the key element of information retrieval process, mixed-language queries have not yet been adequately studied. This study is to determine the user intents of Chinese-English mixed-language queries submitted to a Chinese search engine, and compares the user intents identified by query content to those identified using additional user behavior data (e.g. clicked results, subsequent queries). The preliminary findings present the distributions of user intents by analyzing query only and additional user behavior data, suggesting a specific searching behavior of Chinese-English mixed-language queries users. The findings of this study could provide useful insights in understanding the searching behavior of Chinese-English mixed-language queries users, and enable web search engines to provide users with more relevant results and more precisely targeted sponsored links.

**Keywords:** Information retrieval; multilingual search query; search behavior; user intent; search log analysis
**Citation**: Fu, H., Wu, S. (2015). Determining the User Intent of Chinese-English Mixed Language Queries Based On Search Logs. In *iConference 2015 Proceedings*.
**Copyright**: Copyright is held by the authors.
**Contact**: hf13c@my.fsu.edu, wu1@gmail.com

## 1 Introduction

Although English has remained the predominant content language used in websites, the amount of content in other languages also increases rapidly. In cultures where people use both Chinese and English, using mixed-language in spoken language is very common. In general, a mixed-language consists of terms or sentences mostly in the primary language with certain words in a secondary language. For example, Hong Kong people typically use Cantonese with English words. When such people are doing web searching, they often employ the Chinese-English mixed-language queries to approximate their information need rather than using Chinese-only queries.

A mixed-language query is a search query that is mixed with two or more languages. For example, the query '3G 手机' is a Chinese-English mixed-language query that a user created to look for information about the 3G mobile phones. Researches about Chinese search engines indicated that the number of Chinese-English mixed-language queries has increased in recent years (Chau, Fang, and Yang, 2007). However, only a small number of studies have examined the Chinese-English mixed-language queries in depth (Chau, Lu, Fang, & Yang, 2009). In particular, the Why question (why user performs such search), which is more essential to understand user's information need (Rose & Levinson, 2004), has not been addressed. Identifying user intent of mixed-language queries can provide useful insights in understanding the searching behavior of Chinese-English mixed-language queries users, and can be utilized in improving advertisement targeting and page ranking as well as the presentation of the search results.

Most studies regarding user intent of web searching only analyzed queries. However, queries themselves are not always enough to determine user intent because they are sparse representations of user intents (Silvestri, 2010). Without additional evidence from other sources, such as clicks from search results, think-aloud protocols, and documents viewed, it can be difficult to determine what the users were intending to do in a certain context. User intent of mixed-language query is supposed to be more difficult to infer just by analyzing queries, due to the difficulty of expressing non-specificability need using mixed languages (Belkin, 1980). When user employed a non-native language in the search, the query submitted is more likely to be deviated from the real intent because the user does not know how to use language correctly to form a query. In a previous study, we examined the topic distribution of Chinese-English mixed-language queries and developed a typology of English terms used in those mixed language queries (Fu & Wu, 2014). In this follow-up study, we employed user behavior data such as clicked results, query modification patterns, and subsequent queries to identify user intent, and compared the result to that using query content only.

## 2 Related Work

Several studies have attempted to classify user queries in terms of users' informational actions with different methods, including queries (Broder, 2002; Jansen, Booth, & Spink, 2008, Rose & Levinson, 2004), clicked URL and subsequent queries (Rose & Levinson, 2004), live ticker data (Lewandowski, 2006), and user survey (Broder, 2002; Lewandowski, Drechsler, & Mach, 2012). All these studies followed Broder's taxonomy of query intent: navigational, informational, and transactional. The ratios of user intents varied across different studies and can be found in Table 1. Most studies used English queries except Lewandowski's (2006, 2012).Rose and Levinson (2004) reported that user intents could be inferred with almost no information about the user's behavior. Lewandowski et al. (2012) concluded that the analysis of click-through data is adequate to identify navigational queries but not for other query types.

| Study | Data | Number | Method | N | I | T |
|---|---|---|---|---|---|---|
| Broder (2002)[1] | AltaVista | 3,190 | User survey | 39% | 24.5% | 36% |
| Broder (2002) | AltaVista | 1,000 | Query content | 48% | 20% | 30% |
| Rose & Levinson (2004)[2] | AltaVista | 1,500 (3*500) | Query content, clicked URL and subsequent queries | 61–63% | 11–15% | 21–27% |
| Lewandowski (2006) | Three German search engines | 1,500 | Live ticker data | 45% | 40% | 15% |
| Jansen et al. (2008) | Dogpile.com | 1,523,793 | Query content | 10.2% | 80.6% | 9.2% |
| Lewandowski et al. (2012) | T-Online | 549 | User survey | 34.6% | 38.6% | 26.8% |

Table 1. Comparison of the results from query intent studies using Broder's (2002) taxonomy

## 3 Research Questions

This poster addresses the following questions:

- RQ1: What are the user intents of web searching using the Chinese-English mixed language queries?
- RQ2: Is there any difference in user intents when analyzing subsequent user behavior data beyond the Chinese-English mixed-language queries?

## 4 Data Collection and Research Methods

This study uses queries[3] submitted to the Sogou search engine (http://www.sogou.com/), which was established in 2004 and now is one of the most popular search engines in China. The query-log data was collected in June 2012. Each record consists of six fields: time of click, user ID, query content, ranking of the clicked URL, ordering of user click, and the clicked URL. The query-log data contains 86,538,613 non-empty queries. Query sessions are provided according to cookie information. There are 26,255,952 sessions. One of the researchers developed a C++ program to select and pre-process all the Chinese-English queries that contain both ASCII and double-byte characters along with following user behavior data (clicked results, query modification patterns, subsequent queries) within a session. The final dataset has 346,989 Chinese-English mixed-language queries, accounting for 7.98% of all queries. A random sample of 384 Chinese-English mixed-language queries was drawn from this dataset to conduct content analysis. The sample size was determined using a technique introduced by Powell and Connaway (2004).Table 2 is an example of query with its following user behavior data.

| User behavior | Details |
|---|---|
| Initial query | Maxwell Render 教材 (i.e., Maxwell Render teaching materials) |
| Result click | http://render.haotui.com/forum-16-1.html |
| Result click | http://wenku.baidu.com/view/05e2c21755270722192ef7ed.html |
| Query modify | Maxwell Render v2.6 安装方法 (i.e., Maxwell Render v2.6 installation) |

[1] The ratio of informational and transactional queries was estimated, based on the lower bounds derived from the user survey.

[2] Results are not exact because they aggregate data from their three studies.

[3] The dataset used in this study has been published and can be found at http://www.sogou.com/labs/dl/q-e.html.

| | |
|---|---|
| Result click | http://www.xuanran.net/article/363.html |

Table 2. User behavior following the query *Maxwell Render teaching materials*

This study adopted Rose and Levinson's (2004) approach to manually analyze user intent based on queries and subsequent user behavior. Two researchers independently coded the same set of queries into three different categories of user intents based on query content and user behavior data. When employing user behavior data, if the user modifies the initial query or clicks on more than one result within a session, the last modified query or the last clicked result in this session is used for intent analysis, because if the initial query has been modified or more than one clicks existed, it would imply that the initial query or earlier clicked result(s) have not satisfied the user need completely. Figure 1 and table 3 show the percentages of clicked result numbers for a giving query and query modification patterns in the sample. The researchers achieved an intercoder reliability of 0.872 (Cohen's Kappa) using query content and 0.854 with user behavior data.

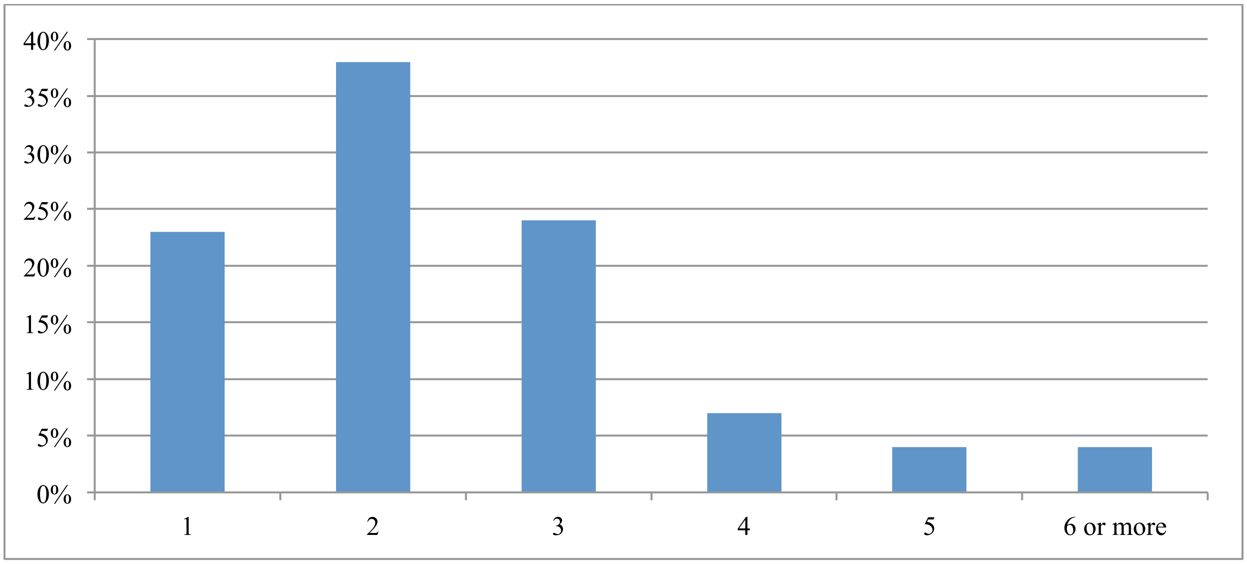

Figure 1. Distribution of the number of clicked results for a given query

| Query modification pattern | Percentage |
|---|---|
| Adding terms | 45.8% |
| Deleting terms | 11.5% |
| Changing the entire query | 37.0% |
| Others | 5.7% |

Table 3. Distribution of query modification patterns within a session

## 5 Preliminary Findings and Discussion

Based on the analysis of query content, of 384 queries in the sample, 52.3% is informational, 39.3% transactional, and 8.3% navigational. The analysis of user behavior data found that 44.3% is informational, 43.2% transactional, and 12.5% navigational. Figure 1 compares our user intent classification with results reported by Broder. While the percentage of informational queries reported in this study is similar, much fewer navigational queries are identified. To be considered navigational, a query should have a single authoritative web site that the user already has in mind (Broder, 2002). The lower percentage of navigational query may suggest users are less certain of if the web pages exist or not, or the effectiveness of mixed-language queries compared to monolingual queries.

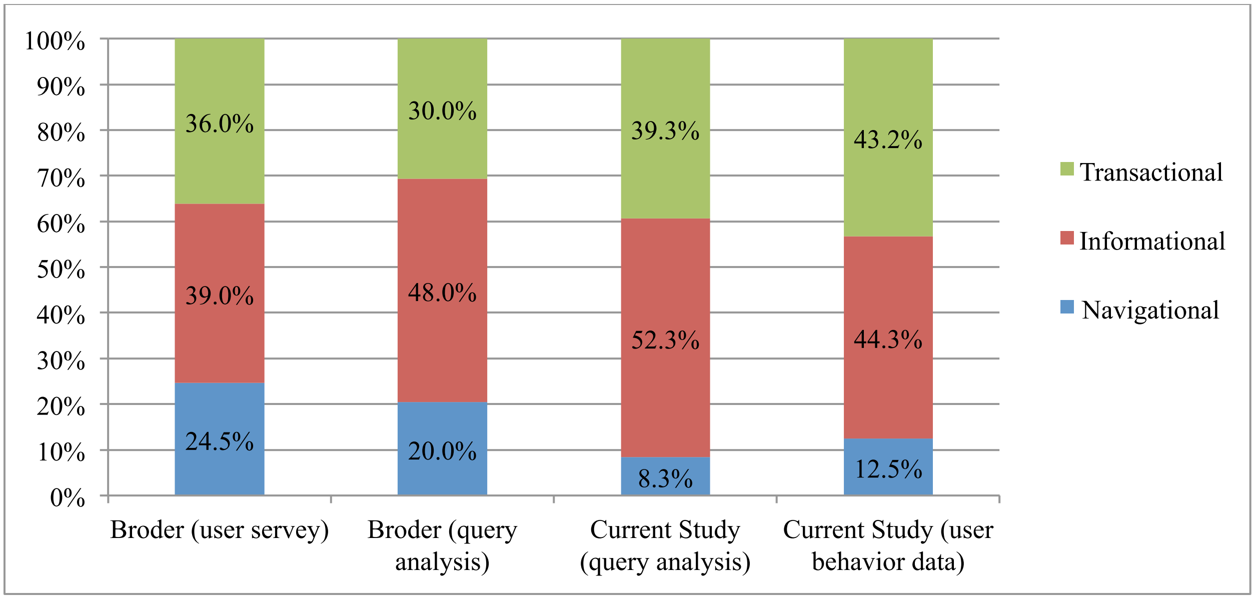

Figure 2. Comparison of Broder's taxonomy to our user intent classification

Rose and Levinson (2004) argued user intents can be detected by using query content only. However, examining additional user behavior data in this study did result in a substantial change in user intents. 26.5% (102 out of 384) queries changed to a different category when analyzing user behavior data. In most cases (86 out of 102), information queries changed to transactional queries, and vice versa. Table 4 shows the distributions of user intent associated with different methods. Chi-square test also indicates these two methods have statistical significances ($X^2$ (2, *N* = 768) = 6.500, p < .05) in classifying user intent. This finding suggests the user intent of mixed-language querying is more difficult to infer. One possible reason is that using non-native language makes a query more ambiguous. Since the English term in a query may not fully represent what is meant by the user, additional information, such as user behavior, is necessary for determining user intent.

| | | | User Intent | | | Total |
|---|---|---|---|---|---|---|
| | | | Navigational | Informational | Transactional | |
| Method | Query Content | Count | 32 | 201 | 151 | 384 |
| | | Expected Count | 40.0 | 185.5 | 158.5 | 384.0 |
| | User Behavior Data | Count | 48 | 170 | 166 | 384 |
| | | Expected Count | 40.0 | 185.5 | 158.5 | 384.0 |

Table 4. Method * User Intent Crosstabulation

## 6 Conclusion and Future Research

This study analyzed the user intents of Chinese-English mixed-language queries submitted to a Chinese web search engine, and compared the user intents identified by query only with those identified by additional user behavior data. The distributions of user intents drawn from queries and using additional user behavior data in this study differ from those of monolingual queries studies, suggesting a specific searching behavior of Chinese-English mixed-language queries users. Future research includes analyzing how and why user intent changes during searching, investigating the context in which users desire to use Chinese-English mixed-language queries, and in which situations these queries would be of beneficial.

## 7 References

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *The Canadian Journal of Information Science, 5*, 133-143.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum, 36*(2), 3-10.

Chau, M., Fang, X., & Yang, C. 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., Lu, Y., Fang, X., & Yang, C. C. (2009). Characteristics of character usage in Chinese web searching. *Information Processing and Management, 45*(1), 115-130.

Fu, H., & Wu, S. (2014). Analyzing Chinese-English mixed language queries in a Web search engine. *Proceedings of the 77th Annual Meeting of the Association for Information Science and Technology*.

Jansen, B. J., Booth, D. L., & Spink, A. (2008). Determining the informational, navigational, and transactional intent of web queries. *Information Processing and Management, 44*(3), 1251-1266.

Lewandowski, D. (2006). Query types and search topics of German web search engine users. *Information Services and Use, 26*(4), 261-269.

Lewandowski, D., Drechsler, J., & Mach, S. (2012). Deriving query intents from web search engine queries. *Journal of the American Society for Information Science and Technology, 63*(9), 1773-1788.

Powell, R. R., & Connaway, L. S. (2004). *Basic research methods for librarians* (4th ed.). Westport, CT: Libraries Unlimited.

Rose, D., & Levinson, D. (2004). Understanding user goals in web search. In S. Feldman, M. Uretsky, M. Najork, & C. Wills (Eds.), *Proceedings of the 13th International Conference on World Wide Web (WWW 2004) (pp. 13–19)*. New York: ACM Press.

Silvestri, F. (2010). Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval, 4*(1—2), 1-174.

## 8 Table of Figures

## 9 Table of Tables