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

Hengyi Fu & Shuheng Wu
School of Information, Florida State University | Graduate School of Library and Information Studies, Queens College, City University of New York

Introduction

- Multilingual information retrieval has emerged as an important research area since the amount of multilingual web resources is increasing rapidly.
- In cultures where people use both Chinese and English, using mixed language in spoken language and web searching is very common. A mixed language query is a search query including words mixed from two or more languages (e.g., Transformers 影评, Python 教程).
- We examined the user intents of Chinese-English mixed language queries using query content and user behavior data (clicked results, query modification patterns, subsequent queries) and compared the results.

Research Questions

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

Data Collection

- Queries submitted to the Sogou web search engine in June 2012, containing 86,539,633 nonempty queries (4,345,557 of them are unique) and 26,255,952 sessions.
- C++ code was developed to select and pre-process all of the Chinese-English queries that contain both ASCII and double-byte characters along with following user behavior data within a session.
- Sampling frame: 346,989 Chinese-English mixed language queries, accounting for 7.98% of all the queries.
- Sample size: 384 Chinese-English mixed language queries along with user behavior data.

<table>
<thead>
<tr>
<th>User Behavior</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial query</td>
<td>Maxwell Render 教材 (i.e., Maxwell Render teaching materials)</td>
</tr>
<tr>
<td>Result click</td>
<td><a href="http://render.haotui.com/forum-16-1.html">http://render.haotui.com/forum-16-1.html</a></td>
</tr>
<tr>
<td>Result click</td>
<td><a href="http://wenku.baidu.com/view/05e2c21755270722192ef7ed.html">http://wenku.baidu.com/view/05e2c21755270722192ef7ed.html</a></td>
</tr>
<tr>
<td>Query modify</td>
<td>Maxwell Render v2.6 安装方法 (i.e., Maxwell Render v2.6 installation)</td>
</tr>
<tr>
<td>Result click</td>
<td><a href="http://www.xuanran.net/article/363.html">http://www.xuanran.net/article/363.html</a></td>
</tr>
</tbody>
</table>

Results and Discussion

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.
- Chi-square test also indicates that these two methods have statistical significances ($X^2 (2, N = 768) = 6.500, p < .05$) in classifying user intent.

The user intent of mixed-language querying is more difficult to infer.

Future Research

Interview users who employ Chinese-English mixed language queries in web searching:
- To investigate the context in which users prefer to use Chinese-English mixed language queries
- To develop a thesaurus of English terms used in MLR to enhance traditional or monolingual controlled vocabularies.

hf13c@my.fsu.edu | shuheng.wu@qc.cuny.edu