Web Query Translation via Web Log Mining*

Rong Hu[1], Weizhu Chen[2], Peng Bai[2], Yansheng Lu[1], Zheng Chen[2], Qiang Yang[3]
[1]Dept. of Computer Science, Huazhong Univ. of Sci. & Tech., Wuhan 430074, China
{ronghu, lys}@mail.hust.edu.cn
[2]Microsoft Research Asia, 5F, Sigma Center, 49 Zhichun Road, Beijing 100080, China
{wzchen, v-pebai, zhengc}@microsoft.com
[3]Dept. of Computer Science, Hong Kong Univ. of Sci. & Tech., Clearwater Bay, Hong Kong
qyang@cse.ust.hk

ABSTRACT
This paper describes a method to automatically acquire query translation pairs by mining web click-through data. The extraction requires no crawling or Chinese words segmentation, and can capture popular translations. Experimental results on a real click-through data show that only 17.4% of the extracted queries are in the dictionary, and our method can achieve 62.2% (in top-1) to 80.0% (in top-5) precision in translating web queries. Moreover, the extracted translations are semantically relevant to the source query, which is particularly useful for Cross-Lingual Information Retrieval (CLIR).

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms
Algorithms, Languages

Keywords
Query translation, click-through log mining, Cross-Lingual IR

1. INTRODUCTION
Query Translation has gained increasing attention in CLIR. Conventional translation methods are mainly based on dictionaries [1], web corpora [2], or domain-specific corpora [3].

In this paper, we propose a novel method to generate query translations based on the analysis of click-through data. Intuitively, millions of users across the world issue queries to a search engine with various languages daily. They formulate queries and click on returned web pages based on their language knowledge, which generates a large-scale and cross-lingual click-through data source. The data thus covers users' understanding of the queries as well as their relations to the clicked pages.

Our main idea of extracting query translation pairs from click-through data is based on two assumptions. The first assumption is that there may exist some naming convention in URLs which specifies the language information of the corresponding pages. To illustrate this assumption, we consider two URLs: http://www.fedex.com/us/, and http://www.fedex.com/cn/. They share the common substring http://www.fedex.com/, and the only difference is the substring indicating the language version, i.e., *us* and *cn* are used to indicate English and Chinese, respectively. We denote two URLs as a bilingual URL pair. The second assumption is that the clicked URLs are relevant to the query. This assumption provides the connection between URLs and queries.

Based on two assumptions, the proposed method consists of two stages: identifying bilingual URL pair patterns, and mining query translation pairs. The method can be used to mine translations in any language pairs, although we currently consider Chinese-to-English query translation pairs in our experiment. To our knowledge, it is the first attempt to leverage click-through data and bilingual URL pairs to mine query translations and demonstrates its effectiveness.

Compared to the previous methods, the proposed method has the following advantages. (1) It can acquire translations for some out of vocabulary (OOV) queries without any need for crawling web pages. (2) It is helpful for CLIR since it can extract semantically relevant queries in target language. (3) It does not require detecting the Chinese phrase boundaries. (4) It captures popular translations of the real queries which are often short without context by exploiting the wisdom of search users.

2. QUERY TRANSLATION EXTRACTION
The identification of URL pair patterns is guided by seed query pairs selected from a bilingual dictionary. Two groups of URLs relevant to these query pairs are found from click-through data (see step 1 in Figure 1). Then the similarity score between the URLs in two groups is computed by edit distance.

The URL pairs whose similarity scores above certain threshold are chosen to extract patterns by the largest common string algorithm (see step 2 in Figure 1). Besides, several simple but very effective filtering techniques are used to improve the quality of the extracted patterns. For example, the patterns containing number are discarded. The difference between URL pairs should not exist in the domain part.

The generated URL patterns are used to guide URL extraction from click-through data. That is, the bilingual URL pairs that correspond to the patterns are selected (see step 3 in Figure 1), which can limit the candidate URL set by only considering the reliable URL pairs.

Once URL pairs are generated, we can find out all queries regarding to the URL pairs based on the second assumption. Then we pair up the bilingual queries and count the frequency of co-occurrence in different URL pairs. The statistically significant candidates are chosen as the final query translation pairs (see step 4 in Figure 1).

*The work was done when the first author and the third author were doing internship at Microsoft Research Asia.

SIGIR'08, July 20-24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

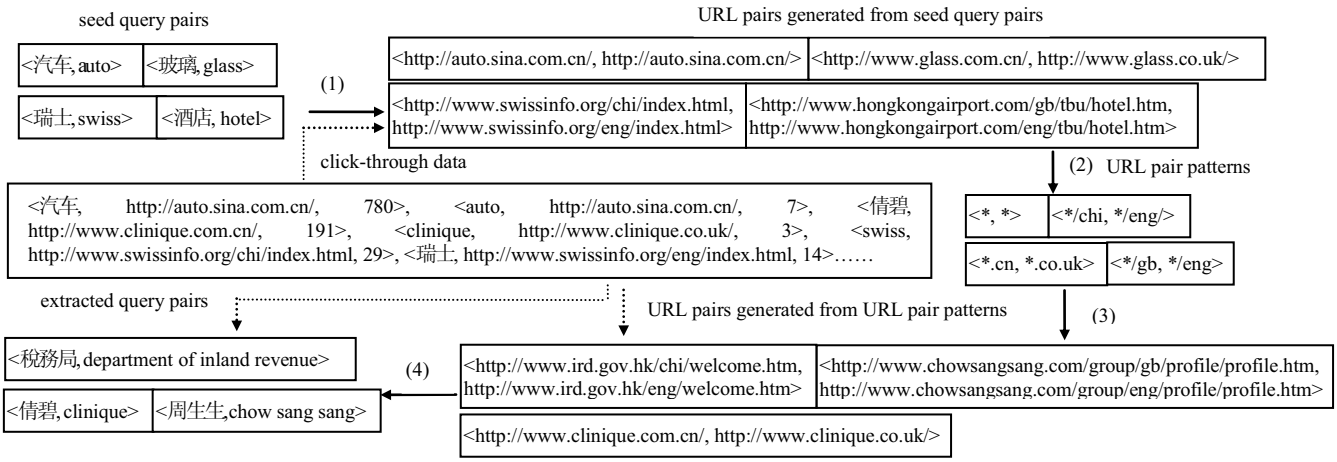

Figure 1: A fragment of data flow using the proposed method

Table 1: The precision values of the extracted query translation pairs

In Dictionary (87 queries)			OOV (413 queries)			All (500 queries)		
Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
63.2%	78.2%	80.1%	62.0%	78.5%	80.0%	62.2%	78.4%	80.0%

Table 2: Examples of the test Chinese web queries and their extracted English translations

	Chinese Queries	Extracted English Translations (Top-5)
In Dictionary	病毒(virus)	virus, blaster, worm, kaspersky, rising
	播放(broadcast)	windows media player, player, media player, media, windows media
OOV	组策略(group policy)	gpmc, gpo, active directory, gmpc, group policy common scenarios
	卡西欧(casio)	casio, casio camera, casio keyboards, casio electronics, caiso camera
	中央芭蕾舞团(national ballet of china)	national ballet of china, national ballet, london national ballet, nutcracker ballet London, giselle ballet

3. EXPERIMENTS

The input of our experiment involves two parts: click-through data, which was collected from a commercial web search engine for eight months spanning from October 2006 until June 2007, and then the sessions containing English and Chinese queries were extracted; and seed queries acquired by LDC dictionary [4].

Due to the high cost of manual judgment in information extraction, we evaluated the experimental results on 500 extracted query translation pairs. Only 87 out of 500 query pairs (17.4%) are included in the LDC dictionary. Our method is found to be effective in translating technical terminologies, named entities, and phrases. The precision of the 87 queries were judged by the LDC dictionary. Human experts were asked to assess the precision of OOV queries. The precision values are shown in Table 1. The results indicate that our method can achieve acceptable results for queries in and out of dictionary.

One advantage of the proposed method is that it can extract relevant translations to benefit CLIR. As shown in Table 2, the extracted top translations are closely related to the source query, even though sometimes they are not the translation equivalent of the source query. So they may help improve CLIR by leveraging the relevant queries frequently used by users.

Another advantage of the proposed method is that it can automatically extract the popular sense of the polysemous queries. For example, the Chinese query "异常" is translated to "abend", "abnormality", "abnormity", "anomalism", or "exceptional", etc in LDC dictionary. The translation in our method is "exception", which corresponds to the information need of most users.

4. CONCLUSIONS

In this paper, we proposed a method to leverage click-through data to extract query translation pairs. The method consists of two stages: identifying bilingual URL pair patterns and matching query translation pairs. Experimental results on a real click-through data show that the method can not only cover 413 the OOV queries out of 500 queries, but also achieve 62.2% (in top-1) to 80.0% (in top-5) precision. Moreover, it can extract semantically relevant query translations to benefit CLIR.

5. REFERENCES

[1] A. Pirkola, T. Hedlund, H. Keskustalo, and K. Järvelin, Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. *Information Retrieval*, 4(3/4), pages 209-230, 2001.

[2] J. Nie, M. Simard, P. Isabelle and R. Durand. Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In *Proceedings of SIGIR'99,* pages 74-81, 1999.

[3] M. Kluck and F. Gey. The Domain-Specific Task of CLEF-Specific Evaluation Strategies in Cross-Language Information Retrieval. *Proceeding of the CLEF 2000 evaluation forum*, 2000.

[4] http://projects.ldc.upenn.edu.