Query Expansion for Personalized Cross-Language Information Retrieval

Dong Zhou[1], Séamus Lawless[2], Jianxun Liu[1], Sanrong Zhang[1], Yu Xu[2]
1. Key Laboratory of Knowledge Processing and Networked Manufacturing & School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, Hunan 411201, China
2. ADAPT Centre, Knowledge and Date Engineering Group, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland
Email: dongzhou1979@hotmail.com

***Abstract*—Cross-language information retrieval research has favored system-centered approaches in the past. The user is not an integral part of the translation and retrieval processes. In this paper, we investigate the problem of personalized cross-language information retrieval by exploiting query expansion techniques. The original query is augmented with terms mined from the user's historical usage information in one language, with the aim of retrieving more relevant results in another language. Experiments semi-automatically constructed by using bilingual Wikipedia documents showed that in general personalized approaches work better than non-personalized approaches. We also found that an individual user model generated from one language can be used to enhance the personalized cross-language information retrieval.**

Keywords—Personalized cross-language information retrieval; query expansion; user model; evaluation

I. Introduction

Cross-language information retrieval (CLIR) is a subfield of Information Retrieval which involves the retrieval of documents in languages that are different to the query's language [1]. It normally requires some facility for language translation incorporated in the process. This is an obvious requirement because query representations and document representations in CLIR systems are not directly comparable. There are three general approaches to translation that can be employed: translate the query representation to match the document representations [2]; translate the document representations to match the query representation [3]; or translate the document and the query representations into a third language or semantic space [4, 5]. There are techniques that use translation resources such as bilingual dictionaries, machine translation systems, and parallel corpora to directly translate the queries and/or documents [2]. Approaches which induce a semantic correspondence between the query and the documents in a cross-language dual space have also been thoroughly studied [4, 6]. However, most studies in CLIR are conducted in a non-user focused manner, although query adaptation and result adaptation have been thoroughly studied.

In this paper we study the personalized CLIR problem through query expansion, whereby the user's original query is augmented by new features with a similar meaning. We adopt various query expansion techniques. These include pseudo-relevance feedback, simple personalized query expansion, penalty query expansion, and techniques based on various methods by computing the similarity between the user query and the user profile. In addition to the comparison of various query expansion techniques, we also want to study the effects of frequency-based user model generation. Another task is to investigate whether user models generated from historical usage information in one language can enhance the search in another language.

II. Related Work

Cross-language information retrieval is a hot and well-studied research area [1]. It normally requires some facility for language translation incorporated in the process. This is an obvious requirement because query representations and document representations in CLIR systems are not directly comparable. There are three general approaches to translation that can be employed: the query representation can be translated to match the document representations [2]; the document representations can be translated to match the query representation [3]; or the document and the query representations can both be translated into a third language or semantic space [4, 5]. Generally, query translation has tended to be favored by the CLIR community, most likely because it is a computationally simpler solution to the mismatch problem. There are techniques that directly translate the queries and/or documents using translation resources such as bilingual dictionaries, machine translation systems, and parallel corpora [2].

Inducing a semantic correspondence between the query and the documents in a cross-language dual space defined by the documents is another commonly adopted approach in CLIR [4, 6]. A technique called latent semantic indexing (LSI) was employed by one of the earliest published CLIR systems [4]. In the study, a comparable corpus was constructed by merging bilingual documents into a document-term matrix. Once this was achieved, a singular value decomposition algorithm could be applied. This process generates a bilingual feature space, which documents and queries could be mapped into. No translation is needed for the CLIR process. LDA-based models

978-1-4673-8395-0/15/$31.00 ©2015 IEEE

are widely used in a similar fashion to help solve the CLIR problem. In particular, Vulić et al.'s work [6], which used a generative bilingual LDA model trained on bilingual Wikipedia documents, demonstrated good performance.

As previously mentioned, most studies in CLIR are conducted in a non-user focused manner, although query adaptation and results adaptation have been thoroughly studied. For example, Cao et al. [7] combined query translation and query expansion in Markov Chains to enhance CLIR. Along similar lines, Ambati and Rohini [8] exploited search logs for cross-lingual query adaption. None of above mentioned work studied the personalized CLIR problem. Moreover, due to the difficulties in user-based study, the experiments conducted in papers related to multilingual IR are in a much smaller scale. In contrast, the current paper presents a comprehensive study on various user profile construction and query expansion techniques as well as an evaluation framework to help lighten the common high barrier in personalized CLIR evaluation.

III. QUERY EXPASION TECHNIQUES

In this section, we present different approaches to perform personalization for CLIR. More specifically, four personalized query expansion techniques are designed. These techniques add terms coming from the user model to the original query.

The first approach is simple query expansion, denoted as **QE**. It works as follows. The first γ keywords in the user model are added to the original query. The user model keywords are ranked in descending order of importance, calculated by techniques defined in Section 3.

The above technique only requires a long query to be performed and is quite efficient. However, the negative effect of the **QE** method is that the expanded query could retrieve documents closer to the user model itself than to the original query. The second technique adjusts the original weight of each user model keyword by using the following equation:

$$Adjust(w) = \delta_1 \cdot \frac{Score(w)}{max_{w'}Score(w')}$$

where δ_1 is the penalty weight. The max () function goes through all user model keywords. After the process, the keywords inside the user model will be re-ordered according to the adjusted weights. Again, the first γ keywords in the user profile will be selected to expand the original query. This technique is denoted as **PQE**, stands for penalty QE.

The next technique exploits co-occurrence of query terms with the user model keywords. Specifically, for each term in the original query, we compute those keywords co-occurring with it most frequently in the user model. This information is used to infer keywords highly correlated with the user query. This algorithm is presented below, denoted as the **CO** technique.

Algorithm 1: Co-occurrence-based query expansion

```
1: Let $S$ be the set of keywords in the user model that could potentially be added as expansion terms to an input query $q$.
2: for each term $t_i$ of $q$ do
    3: $S \leftarrow S \cup Top(t)$ where $Top(t)$ contains the top terms with the closest relationship to $t$ (obtained from co-occurrence statistics)
4: for each term $t_j$ in $S$ do
5: $Score(t_j) = \prod_{t_i \in q}(0.01 + \cos(t_i, t_j))$
6: Select top $\gamma$ terms of $S$ with highest scores.
```

The cosine similarity between two terms t_i and t_j is defined as:

$$cos(t_i, t_j) = \frac{df_{t_i,t_j}}{\sqrt{df_{t_i} \cdot df_{t_j}}}$$

The last query expansion technique employed here is based on the Jaccard coefficient, which is denoted as **JC**. The process is summarized below.

Algorithm 2: Jaccard coefficient-based query expansion

```
1: for each term $t_i$ of $q$ do
2: for each term $t_j$ in the user model do
3: Let $Score(t_j) = J(t_i, t_j)$
4: Select top $\gamma$ terms in the user model with highest scores.
```

The Jaccard coefficient works as follows:

$$J(t_i, t_j) = \frac{|N_{t_i} \cap N_{t_j}|}{|N_{t_i} \cup N_{t_j}|} = \frac{|N_{t_i} \cap N_{t_j}|}{|N_{t_i}| + |N_{t_j}| - |N_{t_i} \cap N_{t_j}|}$$

where $|N_{t_i}|$ denotes the number of documents that contain t_i and $|N_{t_i} \cap N_{t_j}|$ denotes the number of documents that contain both t_i and t_j.

IV. USER MODEL GENERATION

The user model is learned from the user's historical usage information. Here a user is assumed to perform daily searches in one language, and occasionally s/he wants to search information in different languages. So the user model stores terms which represent the user's interests in the main language s/he performs daily searches. A term's weight represents the degree of user's interest in some topics. The information gathering process works as follows: For each query that the user submits, the clicked documents for that query are stored. Note here the query and the documents are in same languages. Then the documents are processed to extract the terms that are most representative in them. To define the representative terms, frequency-based methods can be applied. The extracted terms along with the query terms are subsequently assigned weights accordingly.

As the simplest possible measures, term frequency (*tf*) and inverse document frequency (*idf*) have the advantage of being very fast to compute. The first technique, denote as **TFIDF**, is defined as in the *tf-idf* scheme as follows:

$$Score(w) = \frac{f(t,d)}{max_{t'} f(t',d)} \cdot log \frac{|D|}{df_t}$$

where $f(t,d)$ is the term frequency of term t in document d and the denominator is to prevent a bias towards longer documents. $|D|$ is the total number of documents, df is document frequency.

The next method uses the *BM25* scheme, denoted as **BM25**, and defined as:

$$Score(w) = \sum_t weight_t \frac{(k_1+1)f(t,d)}{K+f(t,d)} \cdot \frac{(k_3+1)f(t,d)}{k_3+f(t,d)}$$

where $weight_t$ is the Robertson/Sparck Jones weight of t and defined as $weight_t = log\frac{|D|-df_t+0.5}{df_t+0.5}$. K is defined as $K = k_1 \cdot ((1-b) + b \cdot \frac{|d|}{avg|d|}$. k_1, b, and k_3 are parameters (set to 1.2, 0.75, and 7 respectively), $|d|$ represents document length and $avg|d|$ stands for average document length.

V. Experiments and Results

In the following section we describe experiments which have been designed to evaluate query expansion and the user model generation techniques. We start the section by discussing the experimental settings, and then we present and analyze the results.

A. Experimental Settings

There is a total lack of personalized CLIR test collections to evaluate the proposed query expansion techniques. So we were obliged to create our own test collection. Motivated by Vicente-López et al. [9], we created a test collection to semi-automatically evaluate the proposed approaches. A Wikipedia database consisting of documents in Chinese and English was used to construct the test collection. Only those articles that are connected via cross-language links between the two Wikipedia databases were selected. A snapshot was obtained on the 14/08/2014, which contained an aligned collection of 158,037 articles in the two languages. The articles are written independently and by different authors, rather than being direct translations of each other. The cross-language information retrieval is performed to use Chinese queries to retrieve English documents. Hence, the Chinese collection is first grouped into 1362 clusters. Each of the clusters could be used to generate a user model. This is possible because in theory clusters should represent different areas of interests. 75% documents inside each cluster are chosen to build the user model, while the remaining 25% are used for testing.

Users are simulated with their user models associated with one or more areas of interests of the document collection (in our case, one cluster represents one particular interest). Here each user is assumed to be interested in the topics of the documents which compose the selected area(s) of interests.

The next question is to generate queries and relevance judgments. This is done by simulating the *cross-language known-item search* [10]. It provides precise semantics and removes the burden to build queries and relevance judgments. The search process assumes that only one document is relevant for a specific query. Here that task is then reduced to find a correct Wikipedia article in the target language with a query provided in the source language. In order to produce more accurate and realistic relevance judgments, we include a human in the loop. 40 undergraduate and postgraduate students manually checked the results retrieved by the training queries. They were instructed to judge each cross-language item as relevant or not by assuming his chosen user model (i.e. the cluster). They were assigned similar numbers of clusters to judge. If s/he feels that most of the documents are not relevant, then that cluster will be discarded. This process filtered out 340 clusters and left 1022 clusters for the final evaluation.

Algorithm 3: Generating queries

```
1: Pick a Wikipedia document $d^E$, find its aligned document $d^C$ in another language
2: Initialize an empty query $q^C$
3: Choose query length $L$ with the Poisson distribution $poi(L)$, the mean is set to the integer closest to the average length of a real web query
4: for each word $w_i$ in $d^C$ do
5:     $Score(w_i) = (1-\delta_2) \cdot P(w_i|d^C) + \delta_2 \cdot P(w_i|Collection^C)$
6: Rank all words from the document $d^C$ based on the scores computed at step 5
7: Select top $L$ words with highest scores to form $q^C$
```

The queries are also automatically generated according to Azzopardi et al.'s work [10]. Formally, suppose there exists a Wikipedia document pair (d^C, d^E), a query q^C will be generated from the document d^C, and then it is used to retrieve the document relevant to q^C, which is implicitly d^E. Since the whole document is too long to be used as a query, the algorithm described in Algorithm 3 is used to generate a much shorter query analog to a real web query. $P(w_i|d^C)$ is calculated as:

$$P(w_i|d^C) = \frac{f(w_i,d^C) \cdot log\frac{|D|}{df_{w_i}}}{\sum_{w_j \in d^C}(f(w_j,d^C) \cdot log\frac{|D|}{df_{w_j}})}$$

We randomly choose 50 users with sufficient training documents. The English terms were processed in the usual way, i.e. down-casing the alphabetic characters, removing the stop words and stemming words using the Porter stemmer. Chinese documents were segmented using a freely available analyzer[1].

[1] http://git.oschina.net/zhzhenqin/paoding-analysis

No other filtering is conducted. Known items in the English collection are assumed to be relevant. Bing Translator[2] is used for translating original and expanded queries. All the information retrieval experiments were performed using the Terrier[3] platform.

In order to usefully evaluate the performance of the user model representation and personalized query expansion methods, 2 different non-personalized runs were selected: **NOP** – a popular and quite robust probabilistic retrieval method using BM25 as the retrieval function, and **PRF** – a pseudo-relevance feedback oriented query expansion method based on the Divergence from Randomness theory.

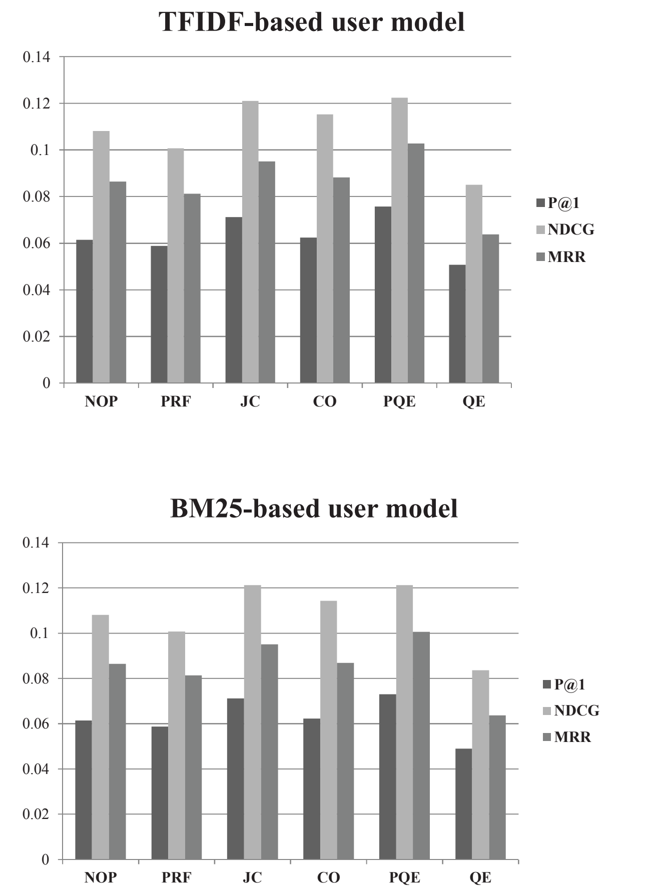
The number of expansion terms and documents for **PRF** is set to 5. γ is empirically set to 5 for all the personalized expansion approaches evaluated here. δ_1 is set to 0.3. According to [10], setting δ_2 to 0.2 efects the average amount of noise within the queries for standard test collections.

B. Results and Discussion

The following evaluation metrics were chosen to measure the effectiveness of the various approaches: mean reciprocal rank (MRR), normalized discounted cumulative gain (NDCG) [11] and the precision of the top 1 documents (P@1). The first two measurements are commonly used to evaluate search algorithms while the last one is useful for evaluating known-item search. The three metrics were calculated for each user and the mean of all the values was calculated, so that the average performance over test users could be computed. Statistically-significant differences in performance were determined using a paired t-test at a confidence level of 95%.

The performance of the techniques used to personalized search, which are shown in Figure 1. As illustrated by the results, the **PRF** model performed consistently poorly for all evaluation metrics except for the **QE** method. This result is not surprising because the evaluation described in this paper is based upon a personalized-approach rather than the non-personalized evaluation model normally employed in large evaluation campaigns. This further demonstrates that using a monolingual user model can enhance personalized CLIR. Pleasingly, except for the **QE** method, all other personalized query expansion-based search models outperform the simpler text retrieval model with the highest improvement of 16.4% inside the chosen group (In terms of the **PQE** method with the MRR metric when compared to **NOP**), which is statistically significant. The low performance of the **QE** method reveals a query-drift problem also found by many other researchers [12]. The expanded queries could retrieve documents closer to the user model rather than the original query. There is a lot of variation across queries in the benefits that can be achieved through personalization. For some queries, everyone who issues the query is looking for the same thing. For other queries, different people want very different results even though they express their need in the same way. Clearly some kind of trade-off is necessary. It has been shown that some queries are better left not personalized. This remains interesting future work.

Fig. 1. Comparison of personalized query expansion techniques



The performance between the **CO** and **JC** methods is quite the same. Both methods put more weight in considering original queries by selecting terms with high co-occurrence statistics with the query terms rather than high weighted user profile terms. However, they are all beaten by the **PQE** method. This further confirms that proper weighting of the profile terms is important in designing any personalized systems.

The result also reveals that both **TFIDF** and **BM25** techniques demonstrated slightly different performance in generating the user model in personalized CLIR. **TFIDF** is consistently better than **BM25** using all personalized query expansion methods. This shows that in terms of user model generation, more complex techniques may not work well in representing the user profile. A possible explanation for this result is that complex techniques are tuned in a much larger corpus rather than the small group of documents inside the user profile. It also confirms that a simple method could yield very good results and is fast to compute.

VI. Conclusions

In this paper, we present a comprehensive study of personalized query expansion techniques for cross-language information retrieval. Experiments showed that in general personalized approaches work better than non-personalized approaches and a user model generated from one language can

[2] http://www.bing.com/translator/
[3] http://www.terrier.org

be used to enhance personalized cross-language information retrieval. In this paper, only one type of personalization strategy has been investigated, i.e. query adaptation. The use of results adaptation and a combination of both approaches will be examined in future research.

Acknowledgment

This research was supported by the National Natural Science Foundation of China under grant numbers 61300129, 61572187 and 61272063, the project Sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, China under grant number [2013] 1792. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

[1] D. Zhou, M. Truran, T. Brailsford, V. Wade, and H. Ashman, "Translation techniques in cross-language information retrieval," *ACM Comput. Surv.,* vol. 45, pp. 1-44, 2012.

[2] J. Gao, J.-Y. Nie, E. Xun, J. Zhang, M. Zhou, and C. Huang, "Improving query translation for cross-language information retrieval using statistical models," presented at the Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, USA, 2001.

[3] D. Oard, "A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval," in *Machine Translation and the Information Soup*. vol. 1529, D. Farwell, L. Gerber, and E. Hovy, Eds., ed: Springer Berlin Heidelberg, 1998, pp. 472-483.

[4] M. Littman, S. Dumais, and T. Landauer, "Automatic Cross-Language Information Retrieval Using Latent Semantic Indexing," in *Cross-Language Information Retrieval*. vol. 2, G. Grefenstette, Ed., ed: Springer US, 1998, pp. 51-62.

[5] T. Gollins and M. Sanderson, "Improving cross language retrieval with triangulated translation," presented at the Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, USA, 2001.

[6] I. Vulić, W. De Smet, and M.-F. Moens, "Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora," *Information Retrieval,* vol. 16, pp. 331-368, 2013/06/01 2013.

[7] G. Cao, J. Gao, J.-Y. Nie, and J. Bai, "Extending query translation to cross-language query expansion with markov chain models," presented at the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, Lisbon, Portugal, 2007.

[8] V. Ambati and U. Rohini, "Using monolingual clickthrough data to build cross-lingual search systems," presented at the New Directions in Multilingual Inforamtion Access Workshop of SIGIR 2006, Seattle, Washington, USA, 2006.

[9] E. Vicente-López, L. de Campos, J. Fernández-Luna, J. Huete, A. Tagua-Jiménez, and C. Tur-Vigil, "An automatic methodology to evaluate personalized information retrieval systems," *User Modeling and User-Adapted Interaction,* pp. 1-37, 2014/06/26 2014.

[10] L. Azzopardi, M. d. Rijke, and K. Balog, "Building simulated queries for known-item topics: an analysis using six european languages," presented at the Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, The Netherlands, 2007.

[11] K Järvelin and J Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.,* vol. 20, pp. 422-446, 2002.

[12] J. Teevan, S. T. Dumais, and D. J. Liebling, "To personalize or not to personalize: modeling queries with variation in user intent," presented at the Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore, Singapore, 2008.