Towards Multilingual User Models for Personalized Multilingual Information Retrieval

M. Rami Ghorab, Dong Zhou, Ben Steichen, and Vincent Wade,

Centre for Next Generation Localisation, Knowledge & Data Engineering Group,
School of Computer Science & Statistics, Trinity College Dublin, Dublin 2, Ireland.
{ghorabm, dong.zhou, ben.steichen, vincent.wade}@scss.tcd.ie

Abstract. The majority of studies in Personalized Information Retrieval (PIR) literature have focused on monolingual IR, and only relatively little work has been done concerning multilingual IR. In this paper we propose a novel method to represent user models in a multilingual fashion. We argue that such representation would be more suitable for Personalized Multilingual Information Retrieval (PMIR). Furthermore, we outline two algorithms for query adaptation based on user information from the multilingual user model.

Keywords: Personalization, Multilingual Information Retrieval, User Modeling, Query Expansion, Pseudo-Relevance Feedback, Web Search.

1 Introduction

Given the enormous amount of information on the web, information can exist in several forms and languages. It may be the case that documents that are relevant to a user's information need exist in languages other than the user's native language. Multilingual Information Retrieval (MIR) is a subfield of Information Retrieval (IR) which involves the retrieval of documents in languages that are different to the query's language [1, 2]. A general characteristic of IR and MIR systems is that if the same query is submitted by different users, the system yields the same results. On the other hand, Adaptive Hypermedia (AH) systems operate in a user-centered manner where services are personalized according to specific user needs [3]. For AH systems to perform personalization, they make use of models to represent user aspects.

Personalized Information Retrieval (PIR) is motivated by the success in the areas of IR and AH [4, 5]. In PIR, different stages of the process are adapted to the user such as the query or the results. The majority of studies in PIR literature have focused on monolingual IR, and only little work has been done concerning multilingual IR.

Our research aims at improving Personalized Multilingual Information Retrieval (PMIR). We investigate how to model different aspects of a multilingual search user, such as preferred language, country, and search interests, and how to exploit this information to personalize the user's multilingual search. In this paper we propose a novel method to represent user models which store the user's search interests as inferred from their multilingual search history. Instead of traditional methods which store terms in a single language [6-8], we propose to store terms in multiple languages. These languages correspond to the original languages of browsed documents, whether the user has viewed them in their original or translated form. The rationale behind our method is to avoid issues related translation inaccuracy if the

user model was monolingual. Furthermore, in this paper we outline two algorithms for query adaptation. The first one performs query expansion using terms from the user model. The second one is a personalized version of Pseudo-Relevance Feedback (PRF) [9-11], where feedback documents are filtered according to the user's interests.

This paper is organized as follows. Section 2 presents related work. Section 3 presents the proposed multilingual user model. Section 4 outlines the proposed algorithms for query expansion. Section 5 outlines the planned experimental settings and the evaluation framework. Finally, Section 6 presents conclusion and future work.

2 Background and Related Work

2.1 Multilingual Information Retrieval

MIR is a subfield of IR that is concerned with retrieving documents from document collections that are not limited to the query's language [2]. In MIR, either the query or the documents can be translated. The query translation approach has gained wider recognition in the literature because of its comparatively lower requirement for resources [12, 13]. A number of studies in the literature aimed to improve retrieval effectiveness in MIR by developing techniques to improve query translation, disambiguation, or adaptation. In [14] the authors propose an algorithm for cross-lingual query suggestion based on multilingual query logs. The authors in [15] propose developing multilingual search systems using bilingual dictionaries and clickthrough information from monolingual search logs. In [16] the authors suggest a Markov Model that combines query translation and expansion in one process.

These studies, in spite of performing adaptation on a multilingual level, do not employ personalization on an individualized scope. We aim at satisfying the specific information needs of a multilingual search user in an individualized manner.

2.2 User Modeling for Personalized Information Retrieval

A key component of PIR systems is the user model. The modeled information is used to personalize the search by adapting the query and/or the results. User information can be obtained implicitly from the search history or explicitly by asking the user to supply information. Different types of representations can be used to represent the user model, mainly: keyword-based, semantic network-based, and concept-based [4].

The authors in [7] implicitly infer the user's interests from browsed web pages. The user model is represented in a keyword-based manner where two vectors are used; one for short-term interests and one for long-term interests. Personalization is employed by re-ranking the search results. The system described in [8] provides a personalized news service, which infers the user's interests from browsing activity. Interest terms are stored in a keyword-based model which is made up of multiple vectors; one for each cluster of interests. Personalization is employed by query expansion and result re-ranking. In [6], a concept-based user model is proposed to represent the user's interests. The model is made up of multiple vectors; one vector per interest category. The categories and concepts are based on the taxonomy of the Open Directory Project (ODP)[1]. Personalization is employed by result re-ranking.

[1] http://www.dmoz.org

The user models in these studies were represented in a single language. We propose to represent user models in a multilingual manner, arguing that this would be more suitable to PMIR since document collections are in different languages. This idea, to the best of our knowledge, has not been previously discussed in the literature.

2.3 Query Adaptation using Pseudo-Relevance Feedback

Query adaptation involves expanding the query with other terms, aiming at retrieving more relevant results [1]. PRF is one of the common techniques used for query expansion [9-11]. It involves performing an initial retrieval round using the source query and implicitly selecting expansion terms from the top N retrieved documents, under the assumption that most of them are relevant to the query. The new query is then submitted to the search engine and the results are presented to the user. The main issue with PRF is that the process is prone to noise caused by the fraction of feedback documents that are not relevant to the query. This may degrade retrieval effectiveness.

The authors in [9] and [11] discuss how automatic classification techniques can be used to identify good and bad terms for query expansion. Several features can be used for the classification process, such as term distribution, term proximity, and document features. In [10], the optimum number of terms for query expansion is investigated.

In the abovementioned studies, feedback terms were filtered with respect to the query and/or the document collection. However, none of these studies took the user into consideration. PIR and PMIR systems seek to retrieve information that is not only relevant to the query, but also relevant to the user. One of the algorithms we propose in this paper uses information about the user's interests to ensure that only feedback terms that are relevant to the user's needs are used for query expansion.

3 A Multilingual User Model

In MIR, documents browsed by the user may be from different languages. Therefore, we propose that for a user model to be more suitable to PMIR systems, it should store terms which represent the user's interests in multiple languages, where a term is stored in the same language of the document from which it was obtained. The proposed user model will maintain a set of weighted-term vectors for each designated language. Multiple vectors in a set represent multiple clusters of interests. A term's weight represents the degree of user's interest in that term within the cluster. Furthermore, each vector is given a weight that indicates the degree of user's interest in the cluster. The proposed user model is an extension to the models in [8] and [6].

When a user clicks on a retrieved result, it may be the case that the document is viewed in its original language or that the system displays a translated version. Considering the latter case, it may seem appropriate to represent the user model in the user's native language only, whereby all terms are translated into that language. However, we believe this would be problematic because: (1) translation inaccuracy may lead to storing terms in the user model which do not represent the user's interests; (2) the translation inaccuracy problem can be exacerbated when the terms in the model are translated, yet another time, for post-translation query expansion; and (3) we should not rule out the possibility that a user may be familiar with multiple languages, and is therefore capable of viewing documents in their original language.

3.1 Populating the User Model

Interest terms will be harvested from the queries submitted by the user and from clicked result documents. The user model will be populated as follows:

1. For each query that the user submits, the clicked documents for that query are grouped by language of the document.
2. For each language group, documents within that group are processed together to extract the terms that most frequently appear in them.
3. The extracted terms along with the query terms (of the original or the translated version of the query) are assigned weights, for example using TF or TF.IDF weighting schemes [1]. The terms and their weights are then stored in a vector.
4. The vector will be added to the user model under the language group that corresponds to the documents from which the terms were extracted, and will be given an initial overall weight.

3.2 Maintaining and Updating the User Model

The number of terms in a vector and the number of vectors to maintain per language will be set to certain thresholds. Following on a modified version of the mechanism in [8], if the maximum number of vectors within a language group was reached, then for any new incoming vector into this group the model will be updated as follows:

1. The incoming vector is added to the other vectors in the group.
2. Using cosine similarity [1], all the vectors in the group are compared to each other. Then, the two most similar vectors are combined by grouping together all terms from both vectors, sorted in descending order of weights. The top terms, according to the threshold, are kept in the combined vector.
3. The weight of the new vector is set as the sum of the weights of the two vectors. Thus, higher vector weights indicate that a vector was subject to merging several times. This may reflect that the topic (cluster) represented by this vector is of high importance to the user as it was repeatedly searched for by the user.

4 Personalized Query Expansion

In this section we propose two alternative algorithms to perform query expansion in a personalized manner. The first algorithm performs query expansion using terms obtained from the user model. The second algorithm performs query expansion using terms obtained via PRF, after filtering out documents that are not relevant to the user.

4.1 Query Expansion Using Terms from User Model

This algorithm aims to adapt the user's query by performing pre-translation and/or post-translation query expansion using terms obtained from the user model. Both kinds of expansion are handled in the same manner; the only difference is the language that is involved in the expansion operation.

In order to expand a query, the vectors of the user model, which belong to the same language of that query, are identified. Then, an important step is to identify which vectors (clusters of interests) from this group are relevant to the topic of the query. Given that the interest vectors are not classified under labeled categories (topics),

identifying relevant vectors can be done in two ways. The first way is to identify the vectors in which the query terms appear. However, this is not sufficient on its own as it may be the case that some of the vectors are relevant to the topic of the query yet they do not have the specific terms of this query. The second way is to perform an iteration of PRF and compare the top N retrieved documents (represented as vectors of weighted keywords) to the vectors in the user model in order to identify candidate relevant vectors to the query. This is performed according to the following steps:

1. For each vector in the user model, calculate the sum of cosine similarities between the vector and each document (represented as a vector).
2. Normalize the sum by averaging over the number of documents, then multiply it by the normalized weight of the vector. Let this be *SimD*.
3. Calculate cosine similarity between the vector and the query. Let this be *SimQ*.
4. Let *SimT* be the total score for the vector, calculated as follows (where k is a constant that controls the influence of query similarity and document similarity):

$$SimT = k\left(\frac{SimQ}{max(SimQ)}\right) + (k-1)\left(\frac{SimD}{max(SimD)}\right) \quad \textbf{(1)}$$

5. After the final score of each user model vector has been calculated, identify the vector that received the highest score. If that vector's score exceeds a certain minimum threshold then select top M terms from that vector and use for expanding the query. Otherwise, do not attempt to expand query (i.e. a low vector score indicates that the vector is not relevant to query's topic, and would therefore degrade retrieval effectiveness if it was used to expand the query).

4.2 Personalized Pseudo-Relevance Feedback

This algorithm is a modified version of the first algorithm, where instead of obtaining expansion terms from the user model, they are obtained from PRF documents. The difference between this algorithm and traditional PRF is that, in this algorithm, the information from the user model is used to select the subset of feedback documents that are most relevant to the user. The terms obtained only from this subset of documents are used for query expansion. This algorithm is expected to reduce the noise caused by irrelevant documents that may appear in feedback documents.

In order to expand a query, the vectors of the user model, which belong to the same language of that query, are identified. Then, these vectors are compared to the PRF documents (represented as vectors) in order to identify the documents which are most relevant to the user. This is performed according to the following steps:

1. For each feedback document, calculate the cosine similarity between that document and each vector in the user model; where each similarity score is multiplied by the normalized weight of the user model vector.
2. Sum across all user model vectors, and then normalize by obtaining the average over the number of vectors in the user model.
3. After all the feedback documents have been scored, sort in descending order of scores and select the top N documents.
4. Analyze the selected feedback documents to extract terms which frequently appear in them, then assign a weight to those terms (e.g. TF, TF.IDF, etc.).

5. Sort the terms in descending order of weights, then select the top M terms and use for expanding the query.

5 Experimentation Outline

5.1 Experimental Setting

The planned experiment will be conducted in a controlled (in-lab) setting, with a group of users from different linguistic backgrounds. The experiment will be carried out on four phases. In the first phase, the users will be asked to use a baseline web search system for their daily search activities over a period of time. The baseline system will be wrapped around one of the major search engines and will not employ any query adaptation. Interactions with the system will be logged. In the second phase, for each user, a subset of the queries and clicked results will be used to build the user model.

In the third phase, the remaining queries will be used for testing. The source (test) queries will be automatically submitted to the search system and the top N results retrieved for each query will be stored in a pool. Furthermore, the system will attempt to re-submit multiple adapted versions of the queries, using the proposed query expansion algorithms and also using traditional PRF, and again, the top N results retrieved for each submission will be pooled with the results of the source queries.

In the fourth phase, the users will be shown the test queries along with the pool of results collected for each one. The users will be asked to provide personal relevance judgements for the results, on a scale from zero to four, where zero indicates that a result is not relevant to their information need, and four indicates that a result is very relevant. All the systems (baseline, traditional PRF, expansion from user model, and personalized PRF expansion) will be evaluated and compared to each other using the Discounted Cumulative Gain metric [1]. This metric is commonly used for calculating IR precision in experiments where documents are judged on a non-binary scale.

5.2 Framework for Experimentation

A framework for PMIR evaluation will be used to carry out the experiments. The framework, which was proposed in [17], is fully implemented in Java and follows the Model-View-Controller architecture. The framework, outlined in Fig. 1, comprises three components: User Modeling, Query Adaptation & Translation, and Result Adaptation & Translation. The *User Modeling* component is concerned with gathering user and usage information about the system users and representing this information in individualized models. In the *Query Adaptation & Translation* component, the user's query is adapted along two stages: pre-translation expansion and post-translation expansion. As discussed in Section 4, query expansion can be based on terms obtained from the user model or from PRF documents. This component makes use of existing state-of-the-art translation techniques for query translation. The output of this component is a set of adapted and translated queries in multiple languages, which are used to retrieve documents from collections in corresponding languages. After document retrieval takes place, the returned result lists are passed to the *Result List Adaptation & Translation* component (one result list in each target language). This component comprises algorithms for result list merging

and re-ranking. This component also makes use of existing techniques to carry out the translation of the title and summary of each result into the user's preferred language.

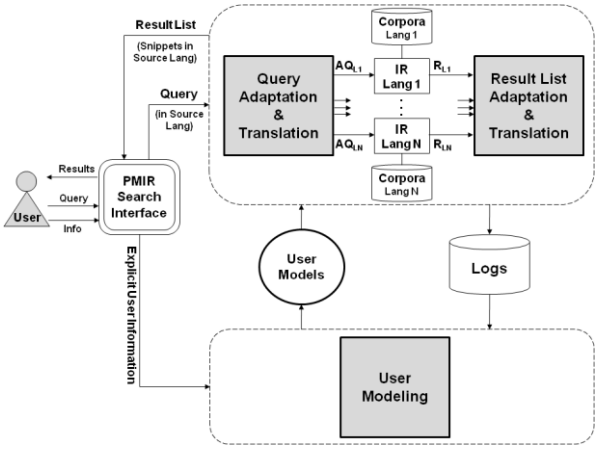

Fig. 1. PMIR Framework

6 Conclusion and Future Work

In this paper, a novel method for representing user models in PMIR systems was proposed. Moreover, an argument was provided regarding why a user model represented in a multilingual fashion may be more suitable to PMIR than a monolingual user model. Furthermore, two algorithms for personalized query adaptation were outlined, and the planned experimental setting was outlined.

After conducting the experiments and performing the evaluation, a viable direction for future work would be to compare query adaptation based on a user model where interest terms are obtained from clicked documents versus a user model where interest terms are obtained from the snippets of the clicked documents.

Acknowledgements

This research is supported by Science Foundation Ireland (grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College, Dublin.

References

1. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
2. Oard, D.W., Diekema, A.R.: Cross-Language Information Retrieval. In: Williams, M. (ed.) Annual Review of Information Science (ARIST), pp. 223-256 (1998)
3. Brusilovsky, P., Millán, E.: User Models for Adaptive Hypermedia and Adaptive Educational Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web, vol. 4321, pp. 3-53. Springer (2007)

4. Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A.: User Profiles for Personalized Information Access. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web, vol. 4321, pp. 54-89. Springer (2007)
5. Micarelli, A., Gasparetti, F., Sciarrone, F., Gauch, S.: Personalized Search on the World Wide Web. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web, vol. 4321, pp. 195-230. Springer (2007)
6. Speretta, M., Gauch, S.: Personalized Search based on User Search Histories. IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005), pp. 622-628, Compiegne University of Technology, France (2005)
7. Sugiyama, K., Hatano, K., Yoshikawa, M.: Adaptive Web Search Based on User Profile Constructed without Any Effort from Users. 13th International Conference on World Wide Web (WWW 2004), pp. 675 - 684. ACM, New York, USA (2004)
8. Chen, L., Sycara, K.: WebMate: A Personal Agent for Browsing and Searching. 2nd International Conference on Autonomous Agents, pp. 132-139. ACM, Minneapolis, Minnesota, United States (1998)
9. Cao, G., Nie, J.-Y., Gao, J., Robertson, S.: Selecting Good Expansion Terms for Pseudo-Relevance Feedback. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008), pp. 243-250. ACM, Singapore, Singapore (2008)
10. Ogilvie, P., Voorhees, E., Callan, J.: On the Number of Terms Used in Automatic Query Expansion. Information Retrieval 12, 666-679 (2009)
11. Leveling, J., Jones, G.J.F.: Classifying and Filtering Blind Feedback Terms to Improve Information Retrieval Effectiveness. Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO 2010), pp. 156-163. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, Paris, France (2010)
12. Ballesteros, L., Croft, W.B.: Resolving Ambiguity for Cross-language Retrieval. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998). ACM, Melbourne, Australia (1998)
13. Zhou, D., Truran, M., Brailsford, T., Ashman, H.: A Hybrid Technique for English-Chinese Cross Language Information Retrieval. ACM Transactions on Asian Language Information Processing (TALIP) 7, (2008)
14. Gao, W., Niu, C., Nie, J.-Y., Zhou, D., Hu, J., Wong, K.-F., Hon, H.-W.: Cross-Lingual Query Suggestion Using Query Logs of Different Languages. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007), pp. 463 - 470. ACM, Amsterdam, The Netherlands (2007)
15. Ambati, V., Uppuluri, R.: Using Monolingual Clickthrough Data to Build Cross-lingual Search Systems. New Directions in Multilingual Information Access Workshop of SIGIR 2006. ACM, Seattle, Washington, USA (2006)
16. Cao, G., Gao, J., Nie, J.-Y., Bai, J.: Extending Query Translation to Cross-Language Query Expansion with Markov Chain Models. 14th ACM International Conference on Information and Knowledge Management (CIKM 2007), pp. 351-360. ACM, Lisbon, Portugal (2007)
17. Ghorab, M.R., Zhou, D., O'Connor, A., Wade, V.: A Framework for Cross-language Search Personalization. 4th International Workshop on Semantic Media Adaptation and Personalization (SMAP 2009), pp. 15-20, San Sebastian, Spain (2009)