A Framework for Cross-language Search Personalization

M. Rami Ghorab Dong Zhou Alexander O'Connor Vincent Wade
Centre for Next Generation Localisation,
Knowledge & Data Engineering Group, Trinity College Dublin,
Dublin 2, Ireland.
{ghorabm, dong.zhou, alex.oconnor, vincent.wade}@cs.tcd.ie

Abstract

Search personalization is an area of considerable research interest. In this paper, we propose a framework for personalizing cross-language search using user models. Our work extends existing studies in two directions. First, the framework extends to the area of cross-language information retrieval. Second, the study aims to elicit features of cross-language search behavior from multilingual search logs. We argue that we can infer a user model, that describes individual user interests and behavior, which can be partially bootstrapped based on choice of interface language. Our experiments involved mining multilingual search logs for interesting patterns of cross-language search behavior. Different patterns were exhibited for users of different languages. The results suggest that there is scope for further investigation on the use of log analysis to improve personalization of cross-language search.

1. Introduction

Over the last decade, much research has been carried out in the literature to improve user satisfaction when using the web. Personalization has become a main principle whereby the user's experience is improved by dynamically customizing a service and presenting it in different manners according to the user.

Search personalization has gained much attention in recent years and many studies were carried out to perform personalization on monolingual search [1, 2, 3, 4]. A key component in these studies is the user model, which represents the user's interests in terms of query and browsing behavior as exhibited in the search logs. The information stored in the model is exploited to adapt the user's query and/or the retrieved results.

Many existing studies aimed at improving cross-language information retrieval (CLIR) by improving the quality of the translation [5, 6, 7, 8]. Other studies attempted CLIR personalization by exploiting search history but without using a user model [9, 10].

The aim of our research is to use user models to improve cross-language search personalization. This research adopts a hybrid approach of Adaptive Hypermedia (AH) [11] and CLIR. In this paper we outline a CLIR personalization framework that comprises three phases: model construction, query adaptation & translation, and result adaptation & translation. We also present initial experimentation for the first phase of the framework. This work is novel in two aspects. First, the framework extends to the area of CLIR, where, to the best of our knowledge, very little work has been done concerning fine-grained personalization. Second, for the model construction phase, in addition to the inclusion of query and browsing behavior, multilingual search logs are analyzed, aiming to elicit multilingual-user-specific attributes and to discover patterns about languages that can help bootstrap the user model and serve as directives for the personalization strategy to follow for users of a certain language.

The rest of this paper is organized as follows: section 2 discusses related work; section 3 provides an illustration of the framework; section 4 describes the experiments; in section 5, results are shown and a discussion of their implications is provided; and finally, section 6 provides conclusion and future work.

2. Related work

2.1. Cross-language information retrieval

CLIR is the subfield of information retrieval that is concerned with retrieving documents that are not limited to the query's language. This enables users to access information beyond their own native language. The two most common approaches for CLIR are either to translate the query into designated target languages, or to translate documents in the collection into

978-0-7695-3894-5/09 $26.00 © 2009 IEEE
DOI 10.1109/SMAP.2009.11

designated languages [6, 7, 8]. The former approach has gained wider recognition in the literature. Many studies argue that the precision of CLIR systems is lower than corresponding monolingual systems and targeted this problem by improving translation and domain disambiguation techniques [5, 10, 9].

Our research will not attempt to improve on translation techniques, but rather, it will make use of existing state-of-the-art techniques for translating the queries and the result snippets. We will instead focus on improvements in personalization.

2.2. User modeling for search personalization

In this section, issues concerned with user modeling are discussed and a review of related work is provided.

A key process in fine-grained personalization is gathering information about the user and representing it in a user model. Three main issues are in consideration:

1. The method by which user information is gathered: either explicitly or implicitly.
2. Content of the model: what information to store in the model and how to represent it.
3. Personalization strategy: whether the adaptation is performed on the query, the results, or both.

Gathering information for user models can be carried out explicitly or implicitly. The former method involves direct user intervention where users explicitly supply feedback and information about themselves to the model. This can be done, for example, by filling feedback questionnaires or web forms that collect demographic data. The explicit method helps gain immediate and specific information about users. However, concerns regarding this method are that users may not wish to exert the extra time or effort to supply the information to the system and that users may sometimes input incorrect or inconsistent information. On the other hand, such information can be gathered implicitly whereby algorithms are used to analyze logs and extract information about the user queries and results viewing behavior. The implicit method has the advantage of not imposing any burden on the user and that models can be updated over time [12].

Several studies in the literature compare the explicit and implicit methods of gathering information. In [13], it is argued that both methods are interchangeable. However, other studies argued that higher accuracy can be reached by combining both methods [14, 15].

Much attention has been given to the implicit method in recent literature. In [1], user behavior data was implicitly incorporated in a user feedback model. The model is made up of features that represent post-search navigation history for pairs of query-URL. Personalization was performed through result re-ranking. In [4], the authors suggest two ways to implicitly build a user profile from search history. The first is based on concepts extracted from the queries, and the second is based on concepts extracted from the snippets of the results that the user viewed. Personalization was performed by result re-ranking. In [3], query disambiguation and query personalization were treated as a unified process of term rewriting. The process used a user profile made up of terms and suggestions of term substitutions that were gathered from search logs. Query history is processed into a term-based graph-form network where the nodes are terms and the edges connecting the terms are suggestions for term expansions. Personalization was achieved through query adaptation. In [2], the authors implicitly construct a user profile based on search history, in addition to a general profile based on concepts extracted from the category hierarchy of the Open Directory Project[1]. The profiles are represented as keyword vectors. Personalization was performed through both, query adaptation and result re-ranking.

From the literature review mentioned above, it can be pointed out that personalization using user models was mainly carried out for monolingual IR. In our work, fine-grained personalization is extended to the field of cross-langue search. As for the content of the user model, it is noted that the majority of studies focus on recording user query and browsing history. A viable extension related to CLIR is the inclusion of a language attribute in the model. Moreover, other user-specific attributes can also be modeled to improve search personalization. As for information gathering, we will combine both, the explicit and the implicit method, with more emphasis on the latter. Finally, it can be seen that the majority of studies perform personalization either by query adaptation or result adaptation alone. In our framework, personalization will be performed on both stages, following on the work reported in [2].

3. Personalization framework

In this section, a personalization framework for CLIR is proposed. The framework, shown in figure 1, comprises three phases: model construction, query adaptation & translation, and result adaptation & translation. An underlying CLIR system ties these phases together and provides the search interface for users. User actions will be logged on the server side.

In the model construction phase, a mixed approach of explicit and implicit information gathering will be followed, with a greater contribution for the implicit

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

method. For the explicit method users will be asked to supply personal information and feedback to the system. For the implicit approach, search history is analyzed by applying various data mining techniques to multilingual search logs. The main data mining technique applied during our initial experimentation with the logs was sequential pattern discovery, which aimed at identifying regular user action sequences.

In addition to the extraction of query and browsing behavior, which was performed by most of the studies covered in section 2.2, we also investigate patterns of behavior that are specifically related to query language.

For this research study, the user model will be represented by a set of feature vectors. The main vector will represent weighted categorical user interests, and for each category, a vector will represent keywords that belong to that category, with weights indicating the degree of user interest in each keyword. The full user model will include features that are drawn from both, the elicited behavior from the logs and additional information obtained explicitly from the user.

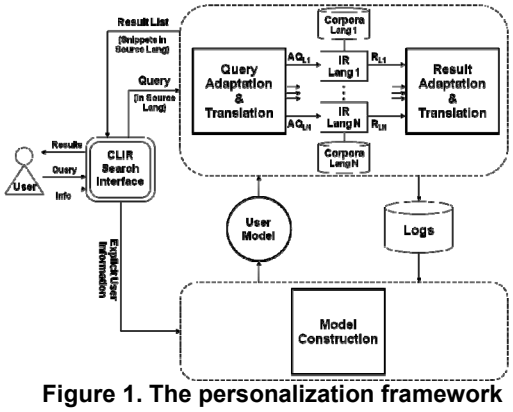

Figure 1. The personalization framework

In the query adaptation & translation phase, the study will look into improving the precision of cross-language search by performing query adaptation using the elicited user model features. The input for this phase is the user's query in its source language. Part of the future experimentation for this phase will investigate whether it is more efficient to adapt the query in its source language or in the target language. The output of this phase is a set of adapted queries that were translated to different languages. Each query will be dispatched to a monolingual IR system, and the results will be collected and passed to the next phase.

In the result adaptation & translation phase, techniques of ranking and merging the results obtained from different language corpora will be investigated. This phase will also use the user model. The output of this phase will be a ranked list of results with snippets translated to the language of the source query.

In order to evaluate the proposed framework, several baseline systems (personalized vs. non-personalized) will be developed to measure the precision and other criteria of the retrieved results.

4. Experimental design

In this section, a discussion of exploratory experiments for the framework is provided. We started with the model construction phase because what is stored in the user model governs what can be done with it. Therefore, the analysis of the multilingual search logs was carried out with the following objectives in mind: (1) to infer user querying behavior with respect to cross-language search; (2) to elicit user-specific attributes of cross-language search users; (3) to identify patterns about languages that can be used to bootstrap the user model when the user specifies a language for his search; and (4) to identify patterns about languages that can serve as directives for the personalization strategy for each language or group of languages.

The experiments were performed on logs recorded for searches done over The European Library (TEL)[2], which is a portal that stands out as a single interface for searching across the content of many European national libraries. The dataset was obtained as part of participation in the Cross Language Evaluation Forum (CLEF 2009)[3]. The logs recorded actions of TEL users along with several fields. The fields are described in the LogCLEF 2009 web page[4]. Our experiments focused on three attributes: *lang* **(**interface language), *action*, and *query*. For the study of actions, the following six actions were considered, as they exhibited a high frequency:

- **search_sim**: simple text box search.
- **search_adv**: advanced search by specific fields of: title, creator (i.e. author, composer, etc), subject, type (e.g. text, image, etc), or language.
- **view_brief**: clicking on a certain library's collection to view its brief list of results.
- **view_full**: clicking on a title link in the list of brief records to expand it.
- **col_set_theme**: specifying a certain collection to search within.
- **col_set_theme_country**: specifying multiple collections for searching or browsing.

It is important to point out the following caveats: (1) the selection of an interface language does not necessarily imply the language of the user query. (2)

[2] http://theeuropeanlibrary.org
[3] http://www.clef-campaign.org/
[4] http://www.uni-hildesheim.de/logclef

The logs did not provide the result that the user viewed.

User session reconstruction was performed by grouping actions with the same session id together in chronological order. Session duration was calculated as the time interval between the timestamp of the first action and the timestamp of last action in the session.

User actions were classified into four categories: *Search* (query actions), *Browse* (browsing/navigating result pages of TEL, excluding following links leading to the browsing of external web sites), *Collection* (actions involving limiting the search scope by the selection of a collection, theme or subject), and *Other*.

5. Results and discussion

5.1. Descriptive statistics

Tables 1 and 2 present descriptive statistics. The logs exhibited outliers, such as the existence of sessions with either a very large number of actions or a single action (max: 1,093; min: 1) and sessions with very long or short durations (max: 116 days; min: 1 second).

Table 1. Frequencies

Item	Frequency
Actions by guests	1,619,587
Actions by logged-in users	12,457
Queries by guests	456,816
Queries by logged-in users	2,973
Sessions	194,627
User IDs	690

Table 2. Central tendencies

Item	Average	Median
Actions per session	8.39	4
Queries per session	2.81	2
Session duration (hh:mm:ss)	00:17:20	00:01:35

It was observed that a small number of actions was performed by logged in users (0.76%) compared to guests (99.34%). Moreover, distinct user ids were found to be just 690. This may indicate that the system does not motivate users to login or that users find it easier, and/or perhaps more secure, not to register with a web search system. Such behavior sets a challenge to fine-grained personalization. Therefore, we need to encourage users to register and login to the system as an explicit way of gathering information about them.

Figure 2 shows the distribution of actions along the broad classification. A significant amount of user actions (11%) were performed before attempting the search, such as the specification of certain collections or subjects to search within. This indicates the diversity of user preferences where many users seek to customize their search environment according to their needs. Recording such pre-search activities in the user model would help towards fine-grained personalization by automatically replicating those preference settings at the beginning of a user session.

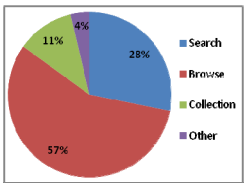

Figure 2. Broad classification of actions

5.2. Interface languages and actions

Table 3. Interface language statistics

Lang.	Number of actions/session		Number of queries/session	
	Average	Median	Average	Median
English	7.97	4	2.7	2
French	9.2	5	3.01	2
Polish	8.63	5	3.14	2
German	9.37	5	3.03	2
Italian	11.3	6	3.73	2

In an attempt to investigate the relation between language and search behavior, several variables were studied across the interface language selected by users of the portal. Recorded actions were distributed along 30 languages. Hereafter, the study focuses on the top five languages in terms of the number of actions. The top language was English (86.47% of the actions), followed by French (3.44%), Polish (2.17%), German (1.48%), and Italian (1.39%). A possible cause for the bias towards English, aside from its inherent popularity, is that it is the default interface language in the portal. Therefore, it may be the case that many non-native English speakers were not aware of the existence of such interface language specification feature in the portal, and were familiar with English sufficiently enough to use it for browsing and navigating the portal. This assumption was supported by the existence of non-English queries associated with actions that were logged under the English language (Caveat lector- however, we cannot rule out the possibility of a native English user searching for a document using its original non-English title). Following our assumption, possible ways to personalize the service and to avoid bias for a default language is to have the system automatically set the language according to a language attribute in the user model or according to the client's IP address.

Table 3 states the average and median for the number of actions and queries per session, and figure 3 shows the frequency distribution of the six main

actions across each of the five languages. It was found that users of English exhibited the lowest average number of actions and queries per session. Moreover, for Italian, the ratio between the number of queries submitted through simple search and those submitted through advanced search was 2.34, while the average ratio for the other four languages combined was 3.51. A probable cause for this kind of inclination may be that queries submitted under Italian were not generally satisfied through simple search, and some users had to reformulate their queries through advanced search. We argue that the modeling of such differences in behavior between users of different languages may serve as a directive for the query adaptation strategy undertaken for a certain language or group of languages.

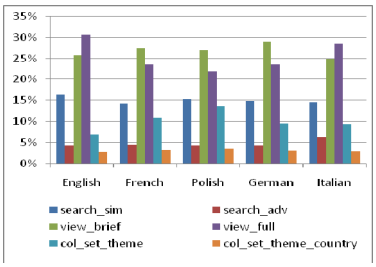

Figure 3. Actions frequency across languages

It was also observed that users of Polish seem to have a higher rate than others in using the feature of specifying a single collection before attempting the search. On the other hand, English was found to have the least rate of usage of this feature. Such observations support the assumptions about the existence of patterns related to languages which can be exploited to improve cross-language search personalization. This suggests that there is scope for further investigation regarding the inclusion of such information about languages and preferred collections in the user model, which can be used, to perform re-ranking on the list of collections.

5.3. Terms frequencies and categories

In the analysis, the number of terms per query and the top queried terms were studied. It was found that the percentage of queries made up of three terms or less was 83.12% in simple search and 69.42% in advanced search. Such trend of users entering fewer search keywords increases the ambiguity of the query, and thus sets challenges for query adaptation.

Table 4, shows the average and median for number of terms per query across interface languages. It was observed that German exhibited the lowest average. Moreover, part of the analysis revealed that German exhibited the largest distribution of queries made up of just one term, while English exhibited the smallest. This may be because the German language allows noun compounds without separating spaces. Such observation indicates how a language may affect the choice of query adaptation strategy.

Table 4. Terms per query

Interface Language	Simple Search		Advanced Search	
	Average	Median	Average	Median
English	2.38	2	3.05	3
French	2.09	1	2.85	2
Polish	1.89	1	2.59	2
German	1.77	1	2.6	2
Italian	2.09	2	3.17	2

As part of the log analysis, the top 20 occurring search terms for each interface language were extracted, excluding stopwords. Furthermore, the terms were divided into five categories: *creator* (author, composer, artist, etc), *location* (cities, countries, etc), *subject* (as per Dewey Decimal Classification), *title* (including proper nouns and common nouns), and *type* (document types, such as: text, image, sound, etc). These categories were mostly based on the fields of the advanced search in the TEL portal, except for location.

It was found that, in simple search, most of the search terms came under the categories of creator and title (30% and 28% respectively). The same was exhibited for advanced search, though with a much greater inclination towards the creator category (45%). This may indicate that user searches were better satisfied by including document creator in the query.

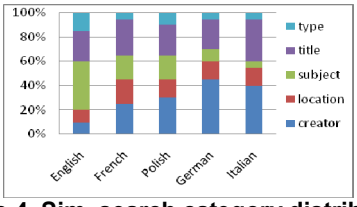

Figure 4. Sim. search category distribution

Figure 4 shows the category distribution of the top 20 search terms across interface languages in simple search. Much difference is observed in search behavior between different languages. For example, in English, 40% of the terms were subjects and 10% were creators, while in German, rather contrasting values were observed where 45% of the terms were creators, and 10% of the terms were subjects. The exploitation of such categorization, and possibly an external knowledge source (e.g. ontology), can be used to disambiguate queries in the query adaptation phase.

6. Conclusion and future work

In this paper we outlined a fine-grained cross-language search personalization framework using user models. The framework comprises three phases: model construction, query adaptation & translation, and result adaptation & translation. A discussion of the experiments for the model construction phase was provided, where a dataset of multilingual search logs was analyzed. The experiments targeted the assumption that we can elicit and augment the user model with attributes that represent: user cross-language querying behavior, multilingual-user-specific features, and language patterns that can be used to bootstrap the user model and serve as directives for the personalization strategy for each language or group of languages.

The results obtained from the analysis support the assumption and thus we conclude that there is scope for further investigation on exploiting search logs to elicit user model attributes for improving personalization of cross-language search. From the analysis, we've seen that we can infer a user model, that describes individual user interests and behavior, which can be partially bootstrapped based on choice of interface language.

In future work we will proceed to establish the full framework then evaluate it for IR performance and user satisfaction. Once this is done, evaluation of the framework will be extended by experimenting with a different dataset of search logs in order to elicit more information and to validate an assumption that an elicited user model for a CLIR system can be transferrable to another CLIR system.

Acknowledgements This research is supported by the Science Foundation of Ireland (grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College, Dublin. The authors would like to thank the reviewers for their valuable comments to the earlier version of this paper.

7. References

[1] E. Agichtein, E. Brill, and S. Dumais, "Improving Web Search Ranking by Incorporating User Behavior Information", in *SIGIR 2006*, Seattle, Washington, USA: ACM, 2006, pp. 19 - 26

[2] F. Liu, C. Yu, and W. Meng, "Personalized Web Search for Improving Retrieval Effectiveness", *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, January 2004.

[3] G. Koutrika and Y. Ioannidis, "A Unified User Profile Framework for Query Disambiguation and Personalization", in *PIA2005*, Edinburgh, Scotland, UK, 2005, pp. 44-53.

[4] M. Speretta and S. Gauch, "Personalized Search based on User Search Histories", in *WI'05*, Compiegne University of Technology, France, 2005, pp. 622- 628.

[5] D. Zhou, M. Truran, T. Brailsford, and H. Ashman, "A Hybrid Technique for English-Chinese Cross Language Information Retrieval", *ACM TALIP*, vol. 7, June 2008.

[6] J. S. McCarley, "Should we Translate the Documents or the Queries in Cross-language Information Retrieval", in *the 37th annual meeting of the Association for Computational Linguistics*, College Park, Maryland, 1999, pp. 208-214.

[7] D. W. Oard and B. J. Dorr, "A Survey of Multilingual Text Retrieval", Technical Report UMIACS-TR-96-19, University of Maryland Institute for Advanced Computer Studies, Maryland, Baltimore, USA, October, 1998.

[8] D. W. Oard, "A comparative study of query and document translation for cross-language information retrieval", in *Third Conference of the Association for Machine Translation in the Americas*, Pennsylvania, USA: Springer-Verlag, 1998, pp. 472-483.

[9] W. Gao, C. Niu, J.-Y. Nie, M. Zhou, J. Hu, K.-F. Wong, and H.-W. Hon, "Cross-Lingual Query Suggestion Using Query Logs of Different Languages," in *SIGIR 2007* Amsterdam, The Netherlands: ACM, 2007, pp. 463 - 470.

[10] V. Ambati and U. Rohini, "Using Monolingual Clickthrough Data to Build Cross-lingual Search Systems", in *New Directions in Multilingual Information Access Workshop of SIGIR,* Seattle, Washington, USA: ACM, 2006.

[11] P. Brusilovsky, "Adaptive Hypermedia", *User Modeling and User-Adapted Interaction,* vol. 11, March 2001.

[12] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli, "User Profiles for Personalized Information Access", in *The Adaptive Web*, 1 ed. vol. 4321, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Germany: Springer-Verlag Berlin Heidelberg, 2007, pp. 54-89

[13] R. W. White, J. M. Jose, and I. Ruthven, "Comparing Explicit and Implicit Feedback Techniques for Web Retrieval", in *TREC-10*, 2001, p. 534.

[14] A. Wærn, "User Involvement in Automatic Filtering: An Experimental Study", *User Modeling and User-Adapted Interaction,* vol. 14, pp. 201-237, 2004.

[15] L. M. Quiroga and J. Mostafa, "Empirical Evaluation of Explicit Versus Implicit Acquisition of User Profiles in Information Filtering Systems", in *Proceedings of the fourth ACM conference on Digital libraries* Berkeley, California, United States: ACM, 1999, pp. 238 - 239.