A study of user profile representation for personalized cross-language information retrieval

Dong Zhou
School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, China
Séamus Lawless
School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland, and
Xuan Wu, Wenyu Zhao and Jianxun Liu
School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, China

Abstract

Purpose – With an increase in the amount of multilingual content on the World Wide Web, users are often striving to access information provided in a language of which they are non-native speakers. The purpose of this paper is to present a comprehensive study of user profile representation techniques and investigate their use in personalized cross-language information retrieval (CLIR) systems through the means of personalized query expansion.

Design/methodology/approach – The user profiles consist of weighted terms computed by using frequency-based methods such as tf-idf and BM25, as well as various latent semantic models trained on monolingual documents and cross-lingual comparable documents. This paper also proposes an automatic evaluation method for comparing various user profile generation techniques and query expansion methods.

Findings – Experimental results suggest that latent semantic-weighted user profile representation techniques are superior to frequency-based methods, and are particularly suitable for users with a sufficient amount of historical data. The study also confirmed that user profiles represented by latent semantic models trained on a cross-lingual level gained better performance than the models trained on a monolingual level.

Originality/value – Previous studies on personalized information retrieval systems have primarily investigated user profiles and personalization strategies on a monolingual level. The effect of utilizing such monolingual profiles for personalized CLIR remains unclear. The current study fills the gap by a comprehensive study of user profile representation for personalized CLIR and a novel personalized CLIR evaluation methodology to ensure repeatable and controlled experiments can be conducted.

Keywords Query expansion, Personalization, Automatic evaluation, Cross-language information retrieval, Topic models, User profile representation

Paper type Research paper

1. Introduction

The World Wide Web is now highly multilingual in nature[1], and users often face the challenge of searching across information in a language of which they are non-native speakers. Cross-language information retrieval (CLIR) has become crucial in addressing this challenge (Nie, 2010; Zhou *et al.*, 2012b). CLIR is a subfield of information retrieval (IR) which involves the retrieval of documents in languages that are different to the query's

Aslib Journal of Information Management
Vol. 68 No. 4, 2016
pp. 448-477

2050-3806
DOI 10.1108/AJIM-06-2015-0091

language. In monolingual IR, users do not always accurately specify their information needs, often using short and ambiguous queries. This problem is exacerbated in CLIR, where users may be able to read information in foreign languages but have limited capacity to formulate suitable queries in that language. It is inevitable that access to the relevant cross-language information is much more difficult than in a monolingual setting.

To overcome the aforementioned query formulation problem in a monolingual environment, researchers have studied personalized techniques (Ghorab *et al.*, 2013). Personalized IR systems do not only retrieve documents that are relevant to a query, but also relevant to a user's individual needs. Such systems greatly help users in satisfying their information needs by minimizing the information overload they experience. In this context, information stored in so-called user profiles can be used to personalize the search process, by altering the initially submitted queries or rearranging the results according to the users' interests and preferences. These user profiles can be stored in an individual manner (Chirita *et al.*, 2007; Xu *et al.*, 2008), or in an aggregate view (Agichtein *et al.*, 2006; Zhou *et al.*, 2014). The user interests are inferred by analyzing the user's search history, extracting keywords from queries that the user submitted and results that the user clicked (Shen *et al.*, 2005; White *et al.*, 2013).

Generally there are two groups of user profile representation techniques: weighted keywords and rich semantic network-based models. The former is made up of terms automatically extracted from documents and information provided by users (Chirita *et al.*, 2007). These keywords are often associated with weights to represent the user's search interests. User profiles can also be represented using a rich semantic network structure (Micarelli and Sciarrone, 2004; Zhou *et al.*, 2012a). In addition to the weighted keywords, the semantic network based user profile can represent weighted relations between semantically related/co-occurring keywords and/or concepts that can accurately depict the user's interests. Sometimes the concepts are not explicitly defined, such as those defined by latent semantic models. However, compared to weighted keyword-based methods, the semantic network-based methods are often computationally complex and hard to use in the latter personalization stage.

User profile representation for monolingual IR has gained significant attention in the literature. However, most of the research in the CLIR or multilingual IR literature is non-user focussed (Zhou *et al.*, 2012b). In the research presented in this paper, we are particularly interested in developing and evaluating IR systems with user profiles. Henceforth, user focussed systems are defined as IR or CLIR systems that utilize user profiles including interactive systems as well as systems with automatically generated user profiles. In comparison, non-user focussed systems are defined as those systems in which no user profiles exist. There have been only a few attempts targeting personalization in multilingual IR (Ghorab *et al.*, 2011, 2012), on a much smaller scale and with simpler methods. Three research questions are raised when representing user profiles in the CLIR process:

RQ1. Can we use the profile representation in one language to enhance cross-language search?

RQ2. Should we take the multilingual dimension of search into consideration when generating user profile representations?

RQ3. Which user profile representation techniques are most suitable for personalized CLIR?

To address the above questions, this paper presents two sets of techniques for user profile representation and studies their effects when used in personalized CLIR.

The first set of techniques extracts the most representative profile terms from a user's historical interactions with the system based on frequency-based methods. The second set of techniques considers latent semantic models to extract profile terms.

With the user profile constructed, personalization can be achieved in a number of ways. In result adaptation, search result lists are re-ranked by incorporating users' profiles accordingly (Cai *et al.*, 2014; Xu *et al.*, 2008). On the other hand, query adaptation attempts to alter the user's initial query by adding, removing or substituting terms selected from the user profiles, with the aim of retrieving more relevant results (Carpineto and Romano, 2012). Personalized query expansion techniques (Chirita *et al.*, 2007; Zhou *et al.*, 2012a), explicit relevance feedback techniques (Ruthven, 2003), and interactive query expansion techniques (White and Marchionini, 2007) are among the popular strategies for query adaptation. To evaluate the profile representations, this paper adopts various query expansion techniques to find suitable methods for personalized CLIR.

In order to evaluate the proposed techniques, suitable collections together with relevance judgments are needed. This process turns out to be very difficult. Although there are many multilingual test collections which have been created by large IR evaluation campaigns like CLEF[2] and NTCIR[3], unfortunately, they are not suitable for evaluating personalized CLIR. The test collections often assume one universal user, and do not provide individual relevance judgments for different users. This situation also applies to monolingual personalized IR. Due to commercial and privacy restrictions, privately owned resources such as e-mails and desktop documents are not easy to acquire. Information stored by a search engine provider is normally unavailable to researchers outside the organization. Researchers have proposed a number of alternative ways to evaluate personalized IR, either through lab-based settings (Teevan *et al.*, 2005), or by using publicly accessible information such as social profiles (Xu *et al.*, 2008; Zhou *et al.*, 2012a). Alternatively, personalized IR can be evaluated using automatically generated test collections. For example, Vicente-López *et al.* (2015) proposed an automatic methodology to evaluate personalized IR systems. However, their method is not quite suitable for evaluating personalized CLIR systems. This is because cross-language relevancy is hard to obtain. In this paper, a novel evaluation method is proposed based on bilingual aligned Wikipedia[4] documents. The procedure of constructing the test collection is inspired by Vicente-Lopez *et al.*'s work, where they use classified documents to simulate users. This procedure will be detailed in Section 5.

The main contributions of this paper are threefold. User profile representation techniques based on frequencies and latent semantics are proposed and systematically evaluated in the context of personalized CLIR. In addition to the user profile construction, various query expansion methods are also introduced and compared. Another contribution is that a novel personalized CLIR evaluation methodology is developed to help lower the common high barrier in personalized CLIR evaluation. This method aims to ensure repeatable and controlled experiments between different personalized strategies, ensuring comparable measures and generalizable conclusions about them.

The rest of this paper is organized as follows. Related work on CLIR, user profile representation, query expansion and latent semantic models is briefly summarized in Section 2. Section 3 describes the user profile representation techniques. Section 4 presents details of personalized query expansion strategies. Section 5 demonstrates our methodology for building a test collection for personalized CLIR, this test collection is semi-automatically generated from Wikipedia to simulate cross-language users.

In Sections 6 and 7 a report is provided on a series of experiments performed to evaluate the user profile representations and personalization strategies. This report includes details of the results obtained. Finally, Section 8 concludes the paper and proposes some future work.

2. Related work

2.1 CLIR

CLIR is a hot and well-studied research area (Nie, 2010; Zhou *et al.*, 2012b). 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 (Gao *et al.*, 2001); the document representations can be translated to match the query representation (Oard, 1998); or the document and the query representations can both be translated into a third language or semantic space (Gollins and Sanderson, 2001; Littman *et al.*, 1998). 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 (Gao *et al.*, 2001). There are also techniques that exploit an intermediate language to translate the source text (Gollins and Sanderson, 2001).

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 (Cimiano *et al.*, 2009; Littman *et al.*, 1998; Sorg and Cimiano, 2008; Vulić *et al.*, 2013). A technique called latent semantic indexing was employed by one of the earliest published CLIR systems (Littman *et al.*, 1998). 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. Latent Dirichlet Allocation (LDA)-based models are widely used in a similar fashion to help solve the CLIR problem. In particular, Vulić *et al.*'s (2013) work, which used a generative bilingual LDA model trained on bilingual Wikipedia documents, demonstrated good performance. Explicit semantic analysis can be used in a CLIR system as an alternative to latent semantic models (Sorg and Cimiano, 2008; Egozi *et al.*, 2011). This technique computes the semantic relatedness between words and indexes documents in relation to a preexisting external knowledge base (e.g. Directory Mozilla [5], formerly known as the Open Directory Project). By using the cross-link references in Wikipedia, this method could be easily extended to CLIR (Cimiano *et al.*, 2009).

Most previous studies in CLIR are conducted in a non-user focussed manner, although query adaptation and results adaptation have been thoroughly studied. For example, researchers performed query expansion for CLIR by using external resources (Lin *et al.*, 2010; Hsu *et al.*, 2008). Cao *et al.* (2007) combined query translation and query expansion in Markov Chains to enhance CLIR. Similarly, Ambati and Rohini (2006) exploited search logs for cross-lingual query adaption. There is also work on results adaptation specifically for CLIR (Zhou *et al.*, 2010). In all of these studies, individual user profiles are not utilized and hence such systems did not provide personalization facilities for CLIR. Steichen *et al.* (2014) presented a survey of polyglot users to analyze

their multilingual proficiency and browsing/search language preferences in personalized multilingual information access. Some attempts have also targeted interactive CLIR in the multimedia domain (Ruiz and Chin, 2010; Vassilakaki *et al.*, 2009). Our previous research investigated multilingual user models to be used in personalized multilingual IR (Ghorab *et al.*, 2011, 2012). However, the user models constructed in this previous research simply considered a vector-based method, and the query expansion method only selected the first few terms from the user models, which is the simplest method used in the current paper. Moreover, due to the difficulties in user-based study, the experiments conducted in those papers are on a much smaller scale. In contrast, the current paper presents a comprehensive study of various user profile construction and query expansion techniques.

With respect to the evaluation of personalized IR, Vicente-López *et al.*'s (2015) proposed a method to be used in monolingual IR. In their method, users are simulated by the areas of interests of documents clustered in the test collection. Each cluster of documents represents a different area of interest or categories (e.g. sports, international, and so on). Queries could be mined from a log file or automatically generated and relevance judgments could also be simulated. In addition, inspired by their work, we propose an evaluation framework to help overcome the common obstacle of personalized CLIR evaluation.

2.2 User profile representation

Monolingual personalized IR systems have been extensively studied in the literature (Ghorab *et al.*, 2013). These systems generally address three main steps: first, information gathering and information representation of user context, which is usually stored as the user profile (Chirita *et al.*, 2007); second, query adaptation and results retrieval, according to the user profile (Zhou *et al.*, 2012a); and third, results adaptation in order to minimize the user effort and maximize user satisfaction (Xu *et al.*, 2008). Among these three steps, the user profile plays a key role in the whole process. Upon gathering information about the user, several techniques and data structures can be used to represent user and usage information. Generally there are two groups of user profile representation techniques: weighted keywords and rich semantic network-based models. The first one is made up of terms automatically extracted from documents and information provided by users (Chirita *et al.*, 2007). These keywords are often associated with weights to represent their search interests. These keywords could also be conceptual/categorical terms that are drawn from some sort of knowledge source (Brusilovsky and Millán, 2007). In this case, external resources should be included in the whole process. Due to the extra overhead, weighted free terms mined from user/usage information are still the most popular approach.

User profiles can also be represented using a rich semantic network structure (Micarelli and Sciarrone, 2004; Zhou *et al.*, 2012a). In addition to weighted keywords, this type of user profile can represent weighted relations between terms and/or concepts that can accurately depict the user's interests. In this case, the user profile uses nodes and associated nodes that capture terms and their semantically related/co-occurring terms, respectively. Weights can be assigned to nodes, associated nodes and the links between them. However, when compared with weighted keyword-based methods, semantic network-based methods are often computationally complex and hard to use in the latter personalization stage.

There are also systems where individual user profiles do not exist. Instead such systems perform search personalization on an aggregate level (Agichtein *et al.*, 2006;

Zhou *et al.*, 2014), where the exploitation of usage information is in a collective manner. It is noted that the use of user profiles has been mostly investigated for monolingual personalized IR systems. However, on the web today, information that is relevant to the user's information need may exist in languages other than the language that the user used to query the system.

The profile representation methods introduced in this paper adopt the weighted keywords approach with consideration of exploring latent semantics to compute the weights. In the approaches outlined above, profile terms in user profiles are normally selected according to their frequencies. In this paper, we provide two additional semantic models to weight potential terms with respect to the latent topics among them. Moreover, to consider the characteristics of CLIR, we use topics obtained from one language to enhance the weights of terms in another language.

2.3 Query expansion

Query expansion attempts to expand or augment the terms of the user query with other terms, with the aim of retrieving more relevant results (Carpineto and Romano, 2012). Pseudo-relevance feedback is a technique which tries to select expansion terms from the top retrieved documents to perform a second round of retrieval (Cao *et al.*, 2008). These terms are assumed to be relevant to the source query. This process is also referred to as local analysis. Global analysis is a technique which implicitly selects expansion terms from a thesaurus, knowledge source, and/or large corpus. Co-occurrence and other statistical measurements are commonly used. These two techniques are non-user focussed. User focussed query expansion techniques can be executed by processing the individual user profile by selecting expansion terms from the profile (Chirita *et al.*, 2007; Zhou *et al.*, 2012a), or by processing the aggregate usage information by obtaining terms from the query logs and/or their associated clicked documents (Agichtein *et al.*, 2006). In practice, acquiring web query logs is difficult for most researchers due to the various concerns of search companies. There are also techniques that require the user to explicitly provide relevance feedback about a number of documents, where expansion terms can be extracted (Ruthven, 2003). Interactive query expansion involves a user interface and allows the user to explicitly select expansion terms from a candidate list of terms suggested by the system (White and Marchionini, 2007). Explicit feedback is often difficult to obtain because users are usually reluctant to participate or have insufficient time to provide such information.

The query expansion used in CLIR is concentrated on pseudo-relevance feedback, with the exception of our previous work described in Ghorab *et al.* (2011, 2012). Either pre- or post-translation expansion can be used (McNamee and Mayfield, 2002). Clearly these techniques are in a non-user focussed manner.

In this paper, we first adopt commonly used query expansion methods in monolingual personalized search systems for the purposes of CLIR. These methods select the top-weighted feedback terms from user profiles for query expansion. In addition, we include methods that select the top-ranked feedback terms that are relevant to the user's current needs for query expansion.

2.4 Latent semantic models

LDA, after it was first introduced in Blei *et al.* (2003), has quickly become one of the most popular probabilistic text modeling techniques and has inspired research in different areas. It has been shown to achieve promising results in modeling text

collections such as news articles (Blei *et al.*, 2003), scientific papers (Wang and Blei, 2011) and blogs (Liu *et al.*, 2009). As a natural extension, various models for multilingual contexts have been proposed. Zhao and Xing (2006) focussed on building latent semantic models suitable for word alignment and statistical machine translation operations. Their models operate on parallel corpora at the sentence level. There have been some efforts that trained on concatenated document pairs in two languages (Littman *et al.*, 1998), but such approaches failed to build a shared latent cross-language topical space. The bilingual LDA model and its extensions (Boyd-Graber and Blei, 2009; Mimno *et al.*, 2009; Ni *et al.*, 2011) train on the individual documents in different languages and their output are joint cross-language topics in an aligned latent cross-language topical space. They have been validated in various cross-language tasks (Vulić *et al.*, 2013) and only require alignments at the document level before training.

Wei and Croft (2006) presented the first large scale evaluation of LDA in monolingual IR, finding it to significantly outperform the query likelihood model. Vulić *et al.* (2013) reported large scale CLIR based on the bilingual model, without the use of any parallel corpora or machine readable dictionaries. However, there are limited approaches to using latent semantic models in constructing the user profile (Zhou *et al.*, 2012a).

In summary, the key gap between current research in CLIR and personalized search is that the majority of studies in the literature investigated personalization in monolingual IR systems, and relatively few studies extended to CLIR. We believe that part of the reasons for this is the difficulties to conduct user-based studies. Moreover, there is an exhibited gap in CLIR literature with respect to performing query expansion based on terms obtained from the user profiles. In the current paper, we try to address these gaps by first introducing and comparing different user profile representation techniques based on frequencies and latent semantics in the context of personalized CLIR. Then we propose various query expansion methods for personalization implementation. Lastly, to overcome the difficulties in personalized CLIR evaluation, we present a novel personalized CLIR evaluation methodology to ensure that repeatable and controlled experiments can be conducted.

3. User profile generation and representation

User profiles are typically learned from a user's usage information. In the scenario outlined in this paper, a user is assumed to perform daily searches in one language, while occasionally s/he wants to search for information in different languages. So the user profile stores terms that represent the user's interests in the primary language that s/he performs daily searches. A term's weight represents the degree of the 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. We assume that the query and the documents are in the same language. It is worth noting that in our paper, users are simulated with their user profiles associated with one or more areas of interest in the document collection. Hence queries and clicked documents are also automatically generated in contrast to studies which use search logs (see Section 5 for details). Then the documents are processed to extract the terms that are most representative of them. To define the representative terms, frequency-based methods and semantic-based methods can be applied. The extracted terms along with the query terms are subsequently assigned weights accordingly. For the bilingual LDA model, documents in different languages are needed. This is illustrated in Section 3.2.

This section presents two sets of techniques for user profile representation and studies their effects when used in personalized CLIR. Both techniques extract and

weight important terms in the queries and documents from a user profile. One set of techniques extracts the most representative profile terms from user's historical interactions with the system and weight the terms based on term frequency (tf) and/or inverted document frequency (idf). These frequency-based techniques are proven to produce good results when used in monolingual personalized IR. The second set of techniques considers latent semantic models to extract and weight profile terms. Both the LDA-based model and the bilingual LDA-based model are detailed here.

3.1 Frequency-based techniques

The first type of user profile generation techniques are based on the frequency with which terms appear in the user's search history. Specifically, the widely used *tf* and *idf* measures are adopted here. They have previously been shown to have a good balance of efficiency and effectiveness (Efthimiadis, 1995). The first technique, denoted as TFIDF, is defined using the *tf-idf* scheme as follows (Baeza-Yates and Ribeiro-Neto, 2011):

$$\text{Score}(w) = \frac{f(t,d)}{\max_{t'} f(t',d)} \cdot \log \frac{|D|}{df_t} \tag{1}$$

where $f(t, d)$ is the term frequency of term t in document d and the denominator is to prevent a bias toward longer documents. $|D|$ is the total number of documents, *df* is document frequency. In other words, the score of a term is highest when the term occurs many times within a small number of documents in the collection. This term thus lends a high level of discrimination to those documents. The score will be lower when the term occurs fewer times in a document, or occurs in many documents in the collection as it then has less power of discrimination. The next method uses the BM25 scheme, denoted as BM25 (Baeza-Yates and Ribeiro-Neto, 2011), and defined as:

$$\text{Score}(w) = \sum_t weight_t \frac{(k_1+1)f(t,d)}{K+f(t,d)} \cdot \frac{(k_3+1)f(t,d)}{k_3+f(t,d)} \tag{2}$$

$$weight_t = \log \frac{|D| - df_t + 0.5}{df_t + 0.5} \tag{3}$$

$$K = k_1 \cdot (1-b). + b \cdot \frac{|d|}{avg|d|} \tag{4}$$

where $weight_t$ is the Robertson/Sparck Jones weight of term t, representing an alternative to inverted document frequency calculation used in the *tf-idf*. 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. This scheme could be viewed as a modification to the tf-idf scheme introduced above.

3.2 Latent semantic models

Another set of user profile generation techniques are based on latent semantic models. In this section the LDA model (Blei *et al.*, 2003) and bilingual LDA model (Boyd-Graber and Blei, 2009; Mimno *et al.*, 2009; Ni *et al.*, 2011) are introduced and their usage in

representing the user profile is outlined. The notations used in the algorithms are listed in Table I:

Algorithm 1. Generative process of the LDA model
1: for each topic $k \in [1, K]$ **do**
2: sample the mixture of words $\varphi \sim Dirichlet(\beta)$
3: for each document $d_j \in [1, M^C]$ **do**
4: sample the mixture of topics $\theta_j \sim Dirichlet(\alpha)$
5: for each word w_i indexed by $i = 1, \ldots, N_{d_j}$ **do**
6: sample the topic index topic $z_{j,i} \sim Mult(\theta_{d_j})$
7: sample the weight of word $w_{j,i} \sim Mult(\varphi_{z_{j,i}})$

The first method to consider is the LDA model (Blei *et al.*, 2003). LDA is a probabilistic monolingual generative model for a text corpus. In contrast to the *tf-idf* and BM25 models discussed above, it is based on the assumption that there exists an unseen structure of "topics" or "themes" in the text corpus, which governs the co-occurrence observations. As such, it claims to discover the latent semantic associations between terms. Consequently, profile terms generated by the LDA model will implicitly incorporate semantic information, and will represent more accurate user profiles.

The generation process in LDA is summarized in Algorithm 1. First, the model generates a mixture of words, grouped into documents (lines 1-2). Then for each document d, a topic distribution is chosen from a Dirichlet distribution (lines 3-4). Next, the word distributions are chosen for each of the topics selected in the previous step. For each word, a particular topic $z_{j,i}$ can be sampled from the document-specific distribution and, finally, a word indicator $w_{j,i}$ is drawn from the topic-specific distribution $\varphi_{z_{j,i}}$ (lines 5-7). In this case, monolinguality is assumed. In particular, it is assumed that the whole process is carried out in language C. The process is repeated for all words in the document. The graphical model corresponding to this process is shown in Figure 1. The darkened circle denotes the variable that is observed and the surrounding plate indicates the independent and identically distributed samples. LDA is a complex model and cannot be solved by exact inference. There are a few approximate inference techniques available in the literature such as variational methods, expectation propagation (Blei *et al.*, 2003; Vulić *et al.*, 2013) and Gibbs sampling. Gibbs sampling is a special case of Markov-Chain Monte Carlo simulation and often yields relatively simple algorithms for approximate inference in high-dimensional models such as LDA. For this reason, we choose to use Gibbs sampling to estimate LDA. In this method, the dimensions of a distribution are sampled alternately

Table I. Notations of LDA model and bilingual LDA model

Symbol	Meaning
M^C	The number of documents in language C
M^E	The number of documents in language E
N_d	The length of document d
K	The number of topics
V	The size of vocabulary
α, β	Dirichlet priors
θ	The topic distributions
φ	The word distributions for language C
ψ	The word distributions for language E

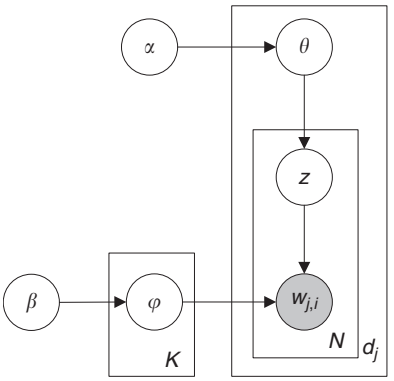

Figure 1. Graphical representation for the LDA model

one at a time, conditioned on the values of all other dimensions. By using Gibbs Sampling, for each word the topic is sampled from:

$$p(z_{j,i} = k) \propto \frac{n_{j,k,\neg i} + \alpha}{n_{j,\cdot,\neg i} + K \cdot \alpha} \cdot \frac{v_{k,w_{j,i},\neg} + \beta}{v_{k,\cdot,\neg} + V \cdot \beta} \tag{5}$$

$n_{j,\,k,\,\neg i}$ counts the number of times that the topic with index k has been sampled from the multinomial distribution specific to document d_j with the current topic $z_{j,\,i}$ not counted. Another counter variable $v_{k,w_{j,i},\neg}$ counts the number of times $w_{j,\,i}$ has been generated by topic k, but not counting the current $w_{j,\,i}$. In this count, a dot $(\cdot)$ denotes summation over all values of the variable whose index the dot takes, that is, all topics in case of $n_{j,\cdot}$ and all words in $v_{k,\cdot}$. By using the above equation, we first initialize the topic assignments z randomly, then in each iteration we sequentially draw the topic assignment of each word. After a predefined number of iterations, we begin recording the final samples. These are posterior distributions calculated as:

$$\theta_{j,k} = \frac{n_{j,k} + \alpha}{\sum_{k'=1}^{K} n_{j,k'} + K \cdot \alpha} \tag{6}$$

$$\varphi_{k,i} = \frac{v_{k,w_i} + \beta}{\sum_{i'=1}^{V} v_{k,w_{i'}} + V \cdot \beta} \tag{7}$$

Algorithm 2. Generative process of the bilingual LDA model

1: for each topic $k \in [1, K]$ **do**
2: sample the mixture of words $\varphi \sim Dirichlet(\beta)$
3: sample the mixture of words $\phi \sim Dirichlet(\beta)$
4: for each document pair $d_j = \{d_j^C \in [1, M^C], d_j^E \in [1, M^E]\}$ **do**
5: sample the mixture of topics $\theta_j \sim Dirichlet(\alpha)$
6: for each word w_i^C indexed by $i = 1, \ldots, N_{d_j^C}$ **do**
7: sample the topic index topic $z_{j,i}^C \sim Mult(\theta_{d_j})$
8: sample the weight of word $w_{j,i}^C \sim Mult(\varphi_{z_{j,i}^C})$
9: for each word w_i^E indexed by $i = 1, \ldots, N_{d_j^E}$ **do**
10: sample the topic index topic $z_{j,i}^E \sim Mult(\theta_{d_j})$
11: sample the weight of word $w_{j,i}^E \sim Mult(\phi_{z_{j,i}^E})$

where θ is the topic distribution for each document j and φ is the term distribution for each topic k. The above two equations counts the number of times each topic assigned to documents and terms from the final read-out samples.

In order to take multilinguality into consideration, a cross-language generative model is further investigated here. This model is named bilingual LDA (Boyd-Graber and Blei, 2009; Mimno *et al.*, 2009; Ni *et al.*, 2011) and is a bilingual extension of the standard monolingual LDA model. The bilingual LDA model has been used in the CLIR process (Vulić *et al.*, 2013), however, it has yet to be exploited in the construction of user profiles. It can model comparable bilingual documents instead of just parallel documents. Bilingual LDA only requires bilingual documents to be aligned at the document level, rather than at sentence level and/or word level. That is, two bilingual documents need not to be a sentence-by-sentence/word-by-word translation of each other. As long as they discuss at least a portion of similar themes, the bilingual LDA model can effectively learn the cross-lingual topics where words are closely semantically related. There is also no restriction on document length for documents written in two languages. The model learns topics which are shared between the two languages. These topics, and the word distribution over them, will be employed to construct user profiles for use in personalized CLIR. In theory, the user profiles represented by the bilingual LDA model should be more comprehensive than the monolingual LDA model. This is confirmed by our experiments reported later in the paper.

Unlike the standard LDA model, the bilingual LDA model uses θ to represent the topic distribution of documents in two languages. As stated above, it should be noted that there is a comparable corpus aligned at the document level, instead of sentence level and/or word level. Therefore, θ can be viewed as a language independent factor, and shared among comparable bilingual aligned documents. The generation process for the bilingual LDA model is slightly different from the LDA model. This process is summarized in Algorithm 2. Now we start to sample the topic distribution θ_j for each document pair d_j instead of for one monolingual document (lines 4-5). Then, a topic $z^C_{j,i}$ is sampled with respect to θ_j for language C. A word $w^C_{j,i}$ in the document of the current document pair d_j is then generated from a multinomial distribution $\varphi_{z^C_{j,i}}$ (lines 6-8). At the same time, a word $w^E_{j,i}$ of the language E is also sampled following the same procedure (lines 9-11). It is worth noting that the words at the same positions for paired documents in different languages need not be from the same cross-language topic. The only constraint is that the overall distributions of topics over documents in a document pair modeled by θ_j are the same. The graphical model corresponding to this process is shown in Figure 2.

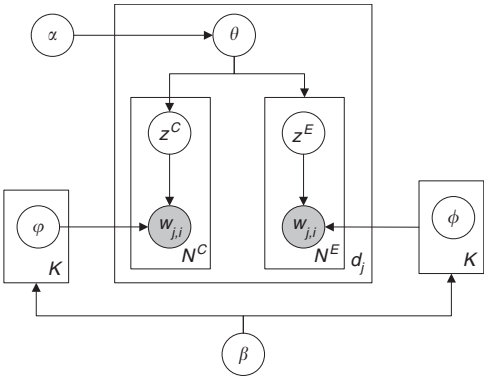

Figure 2.
Graphical representation for the bilingual LDA model

Similar to the LDA model, the updating formulas for Gibbs sampling for bilingual LDA are:

$$p\left(z_{j,i}^{C}=k\right) \propto \frac{n_{j,k,\neg i}^{C}+n_{j,k}^{E}+\alpha}{n_{j,\cdot,\neg i}^{C}+n_{j,\cdot}^{E}+K\cdot\alpha}\cdot\frac{v_{k,w_{j,i}^{C},\neg}^{C}+\beta}{v_{k,\cdot,\neg}^{C}+V^{C}\cdot\beta} \tag{8}$$

$$p\left(z_{j,i}^{E}=k\right) \propto \frac{n_{j,k,\neg i}^{E}+n_{j,k}^{C}+\alpha}{n_{j,\cdot,\neg i}^{E}+n_{j,\cdot}^{C}+K\cdot\alpha}\cdot\frac{v_{k,w_{j,i}^{E},\neg}^{E}+\beta}{v_{k,\cdot,\neg}^{E}+V^{E}\cdot\beta} \tag{9}$$

The meaning of the symbols used is the same as in the LDA model, except this time for dual languages C and E. $n_{j,k,\neg i}$ counts the number of times that the topic with index k has been sampled from the multinomial distribution specific to document d_j with the current topic $z_{j,i}$ not counted. Another counter variable $v_{k,w_{j,i},\neg}$ counts the number of times $w_{j,i}$ has been generated by topic k, but not counting the current $w_{j,i}$.

Having outlined the basic LDA model and the bilingual LDA model, the techniques for representing a user profile based on these two models are presented. The third technique, which is denoted as LDA-1, is defined as:

$$\text{Score}(w_i)=\sum_{k=1}^{K}\sum_{j=1}^{M}\varphi_{k,i}\cdot\theta_{j,k} \tag{10}$$

where θ and φ are introduced as above. We sum the product of the topic distribution and term distribution over all documents and across all topics to obtain the final weights for a particular term.

To compare the different strategies for calculating the weight of term w, the fourth technique simply uses $\text{Score}(w_i)=\sum_{k=1}^{K}\varphi_{k,i}$. This method only uses the term distribution and is denoted as LDA-2.

For the bilingual LDA model, a similar scoring process could be defined. The only problem is that a document-aligned corpus should be used to train the model. This could be done by simply translating the documents stored in the user profile to the languages that the user wants to search for by using automatic approaches such as machine translation. Alternatively, if the test collection is constructed by using comparable corpora such as the experiments described in this paper, every document in the user profile will already have its bilingual counterpart. See Section 5 for details about this process. Given the document-aligned bilingual usage information, the weight of a particular term is defined as:

$$\text{Score}\left(w_i^{C}\right)=\sum_{k=1}^{K}\sum_{j=1}^{M}\varphi_{k,i}\cdot\theta_{j,k} \tag{11}$$

Again we sum the product of the topic distribution and term distribution over all documents and across all topics for terms in the user's native language. Here the user's native language is assumed to be language C. This technique is denoted as BLDA. At the moment the statistics calculated in language E are not used because the

document in that language is just employed to produce a more accurate model than by using the monolingual LDA model. Integrating the documents in language E into the model remains future work.

Here we do not directly use the latent topics produced by the latent semantic models because we found that they could not improve the performance of personalized CLIR through query expansion. However, we utilize these topics when calculating the weights for profiles terms, rather than generating weights using term frequencies alone. Experimental evaluation confirms the usefulness of these methods.

4. Personalized query expansion

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

Algorithm 3. Query Expansion procedure based on co-occurrence statistics
1: Let S be the set of keywords in the user profile 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 top terms with the closest relationship to query terms (obtained from co-occurrence statistics)
4: **for** each term t_j in S **do**
5: $Score(t_j) \leftarrow \prod_{t_i \in q}(0.01 + \cos(t_i, t_j))$
6: Select top γ terms of S with highest scores.

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

The above technique only requires the system to perform a long query and is quite efficient. The negative effect of the QE method is that the expanded query could retrieve documents closer to the user profile itself than to the original query, while the original query represents the user's current information need. Moreover, the added profile terms could also retrieve more irrelevant documents. Both problems will become more pronounced as more profile terms are added. However, adding too few terms may cause a poor representation of the true preferences of the user, so we need a trade-off here.

So the second technique adjusts the original weight of the profile keyword by adding a global normalization factor applied to the weights of the profile terms:

$$\text{Adjust}(w) = \delta_1 \cdot \frac{\text{Score}(w)}{\max_{w'} \text{Score}(w')} \tag{12}$$

where δ_1 is another free parameter to tune. The max() function goes through all profile keywords. After the process, the added profile terms can receive at most a fraction δ_1 of the maximum weight attached to the original query terms. The normalization factor 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 profile keywords (Chirita *et al.*, 2007). Specifically, for each term in the original query, the technique

computes those keywords co-occurring with it most frequently in the user profile. Then this information is used in order to infer keywords highly correlated with the user query. This co-occurrence-based query expansion algorithm is presented in Algorithm 3, denoted as the CO technique. In line 5, we add 0.01 to avoid zero values. The cosine similarity between two terms t_i and t_j is defined as:

$$\cos\left(t_i, t_j\right) = \frac{df_{t_i,t_j}}{\sqrt{df_{t_i} \cdot df_{t_j}}} \tag{13}$$

Algorithm 4. Query expansion procedure based on Jaccard coefficient
1: **for** each term t_i of q **do**
2: **for** each term t_j in the user profile **do**
3: Let $Score(t_j) \leftarrow J(t_i,t_j)$
4: Select top γ terms in the user profile with highest scores.

The last query expansion technique employed here is based on the Jaccard coefficient, which is denoted as JC. The process is summarized in Algorithm 4, and the Jaccard coefficient works as follows:

$$J\left(t_i, t_j\right) = \frac{\left|N_{t_i} \cap N_{t_j}\right|}{\left|N_{t_i} \cup N_{t_j}\right|} = \frac{\left|N_{t_i} \cap N_{t_j}\right|}{\left|N_{t_i}\right| + \left|N_{t_j}\right| - \left|N_{t_i} \cap N_{t_j}\right|} \tag{14}$$

where $\left|N_{t_i}\right|$ denotes the number of documents containing t_i and $\left|N_{t_i} \cap N_{t_j}\right|$ denotes the number of documents containing both t_i and t_j. The JC method is slightly different from the CO method. JC does not include the additional $Top(t)$ terms as we found that including these terms decreases the performance of the method. This concludes the discussion of the proposed personalized query expansion techniques. In the next section, a framework to evaluate personalized CLIR is presented.

5. Personalized CLIR evaluation framework

5.1 Monolingual personalized IR evaluation framework

The evaluation framework described here is inspired by Vicente-López *et al.*'s recent introduction of ASPIRE, an automatic evaluation methodology for personalized IR systems (Vicente-López *et al.*, 2015). The method joins the advantages of both system-centered and user-centered evaluation approaches, aiming to provide repeatable, comparable and generalizable test collections for personalized IR. The authors compared the results produced by the ASPIRE framework with those obtained from a user study. Vicente-López *et al.* suggested that this framework may be considered as an interesting alternative to costly and difficult user studies. The results obtained using the framework are very close to the real normalized discounted cumulative gain (NDCG; see Järvelin and Kekäläinen, 2002) values obtained from the user study (Vicente-López *et al.*, 2015, p. 25). It is able to discriminate between either different personalization techniques or different parameter configurations of a given personalization method (Vicente-López *et al.*, 2015, p. 25). Furthermore, it is able to discriminate differences in performance between the different user profile configurations of a given personalization method (Vicente-López *et al.*, 2015, p. 26). As the main goal in this paper is to select the best user profile representation

techniques and the best personalization strategies for personalized CLIR, this makes it an ideal tool to use here.

Algorithm 5. Query generation from Wikipedia document
1: Pick a Wikipedia document d^E, find its aligned document d^C in another language
2: Initialize an empty query without terms 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: $Score(w_i) \leftarrow (1-\delta_2) \cdot P(w_i|d^C) + \delta_2 \cdot P(w_i|Collection^C)$
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

In Vicente-López *et al.*'s evaluation framework, there are four main components specified. Document collection should be classified into different areas of interest or categories. This can be achieved using explicit clustering and/or implicit categorization. Users are simulated with their user profiles associated with one or more areas of interest in the document collection. Here each user is assumed to be interested in the topics of the documents which compose the selected area(s) of interest. Queries can be formulated by real users and relevance judgments are determined by a simulation approach. If a document belongs to the area of interest of a particular user profile, it will be considered as relevant to the given user profile.

However, Vicente-López *et al.*'s method could not be directly used in personalized CLIR, as cross-language relevancy is more difficult to obtain. This is a non-trivial task. In the following section we extend the ASPIRE framework to be used in personalized CLIR.

5.2 Evaluation framework for personalized CLIR

In order to find suitable document collections for evaluating personalized CLIR, bilingual aligned documents are needed. In such cases, Wikipedia serves as an ideal collection. A very important characteristic of Wikipedia articles is that they are actually linked across languages. Cross-language links are those that link a certain article to a corresponding article in the Wikipedia database in another language. They discuss the same topic, but vary in style, length and vocabulary, and are authored by different users. They share a certain number of main concepts across languages.

To simulate users, Wikipedia articles are first clustered into several categories to represent user interests. This is done in the source language, which is assumed to be user's daily search language. Any clustering algorithm can be used here. In the experiments below, *K*-means is used to produce hard clusters. User profile representations are then computed inside these clustered documents. We use one cluster to represent a user's profile. Documents inside a cluster may discuss a variety of themes, which represent a user's diverse interests. In this research we assume that a user has consistent interests within a certain topic. If a user has multiple interests across different clusters, this will pose a problem. How to deal with this scenario remains as future work.

It is then necessary to generate queries and relevance judgments. This is achieved by simulating cross-language known-item search (Azzopardi *et al.*, 2007). Given a Wikipedia document in one language as a query, the process presumes only that the corresponding document in another language is relevant. So the relevant documents could be viewed as relevance judgments performed by Wikipedia authors. As a whole

document is typically too long to be used as a meaningful query, an automatic process is utilized according to Azzopardi *et al.*'s work. In the user profile generated at the last step, 75 percent of documents are kept as training documents to represent the user's interests. The remaining 25 percent documents are used to generate queries. 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 5 is used to generate a much shorter query analog to the real web queries. $P(w_i|d^C)$ is calculated as follows:

$$P\left(w_i|d^C\right) = \frac{f\left(w_i, d^C\right) \cdot log\frac{|D|}{df_{w_i}}}{\sum_{w_j \in d^C}\left(f\left(w_j, d^C\right) \cdot log\frac{|D|}{df_{w_j}}\right)} \tag{15}$$

The probability of a term being selected is proportional to the tf/idf of the term in the document. The strategy is to improve the discrimination of query terms selected. In this way, terms that are highly discriminative within the collection, and also very popular within the document are more likely to be selected (see further Azzopardi *et al.*, 2007).

6. Experimental settings

In the following section a series of experiments are described which have been designed to evaluate the user profile representation and query expansion techniques described above. This evaluation focusses on the following thematically related questions:

RQ4. Does the user profile generated from one language enhance personalized CLIR?

RQ5. Are the personalized query expansion methods an improvement over classical non-personalized methods when used in personalized CLIR, and which method is the most effective?

RQ6. Will the user profile built upon bilingual aligned documents prove an advance over the single language-based profile?

RQ7. Which user profile representation techniques are most suitable for personalized CLIR?

RQ6 and *RQ7* here are the same as the *RQ1* and *RQ3* introduced at the beginning of this paper. However, *RQ1* from the introduction section has been divided into *RQ4* and *RQ5* here to separately investigate the user profile generation (*RQ4*) and personalization strategies (*RQ5*).

6.1 Experimental data

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 August 14, 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 CLIR is performed by first translating Chinese queries into English queries using a translation mechanism[6], then retrieving English documents from the test collection. Hence, the Chinese collection is first grouped into 1,362 clusters. Each of the clusters could be used to generate a user profile.

This is possible because, in theory, clusters should represent different areas of interests. In total, 75 percent of documents inside each cluster are chosen to build user profiles, while the remaining 25 percent are left for testing. Note, the number of clusters is not fixed to a value as when clustering the category of the document is a lso considered.

In order to produce more accurate and realistic relevance judgments, one further step was taken. In total, 40 undergraduate and postgraduate students were hired to manually check the results retrieved by the training queries. They were instructed to judge whether each cross-language item was relevant to the given query (usually two or three words long) or not by assuming his chosen user profile (the top keywords extracted from a cluster). This relevancy is quite easy to judge as in theory only one item is relevant to a given query. A simple system was developed to perform the task. The subjects were assigned similar numbers of clusters to judge. If s/he felt that most of the given queries could not generate relevant items, or the relevant items were not consistent with the user profile, then that cluster was marked. Each cluster was reviewed by at least three subjects, and if two or more mark the cluster, then it was discarded. This process filtered out 340 clusters and left 1,022 clusters for the final evaluation. We manually checked the relevance judgments for two reasons. First, possible errors produced by the query generation process. Sometimes the corresponding article may not be relevant to a generated query. Second, even if the article is relevant, it may be not relevant to the user profile generated.

In the experiment, we wanted to fully study different kinds of users. Four groups of users were simulated according to the number of training documents associated with clusters: users with less than 50 documents, denoted as WIKI50; users with 50-100 documents, denoted as WIKI100; users with 100-500 documents, denoted as WIKI500 and finally users with more than 500 documents, denoted as WIKIgt500. This choice reflects users who have rich information in the user profiles as well as those who have less rich information in the user profile. This is consistent with previous research concerning real-user studies (Xu *et al.*, 2008; Zhou *et al.*, 2012a). Users with less than 50 documents represent less active users within the system, where the user profile may only contain limited information for personalization. Users with 50-100 and 100-500 documents can be viewed as normal active users while users with more than 500 documents are very active users. For those users, rich user profiles can be built. In total, 50 randomly selected users from each group were extracted to form a total collection of 200 test users. The English terms were processed in the usual way: 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[7]. No other filtering was conducted. Queries were generated by following the procedure discussed in Section 5.2. Known items in the English collection were assumed to be relevant. Bing Translator[8] was used for translating original and expanded queries. All the IR experiments were performed using the Terrier[9] open source platform.

6.2 Evaluation methodology

The following evaluation metrics were chosen to measure the effectiveness of the various approaches: mean reciprocal rank (MRR), the NDCG and the precision of the top one 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 query 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 percent.

6.3 Experimental runs

In order to usefully evaluate the performance of the user profile representation and personalized query expansion methods, two 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. This approach has previously shown good results (Amati and Rijsbergen, 2002), which is also a natural choice for evaluating the difference between expanding queries by selecting the terms from the user profiles and from relevant documents[10].

6.4 Parameter setting

A subset of simulated users (100 user profiles) was also randomly selected to train the necessary parameters described below. There was no overlap between the training-set of simulated users and the test-set of simulated users.

The selection of the number of expansion documents and the number of expansion terms for PRF is illustrated in Figure 3(a). As the results suggested, fewer numbers tend to work well. So the number of expansion terms and documents is set to 5. Similar effects can be observed in Figure 3(b) for the number of expansion terms (the variable γ) for the TFIDF user profile representation by trying various personalized expansion methods. Hence, γ is set to 5 for all the personalized expansion approaches evaluated here. In order to fix the parameters δ_1, a number of runs were executed with a spread of settings from 0.1 to 0.9. As shown in Figure 4, the best results were obtained when the value is 0.3. According to Azzopardi *et al.* (2007), setting δ_2 to 0.2 effects the average amount of noise within the queries for standard test collections.

The parameters for the LDA and bilingual LDA models are set as follows. α and β are set to $50/K$ and 0.01, respectively, where K is set to a third of the size of the user profile tested. The number of iterations for both models is set to 1,000.

7. Experimental results

In this section experimental results are presented and discussed. The frequency based user profile representation techniques are first compared to the latent semantic techniques. Then the performance of various personalized query expansion methods is reported. Finally, the section illustrates the performance of the proposed personalization strategies for CLIR across different groups of users.

7.1 Comparison of user profile representation techniques

This set of experimental results describes the performance of the techniques used to represent user profiles, which are shown in Table II. Statistically significant results with respect to the NOP approach are marked in italics.

7.1.1 Performance of frequency-based techniques. Table II reveals that both TFIDF and BM25 techniques demonstrated slightly different performance in representing user profiles in personalized CLIR. TFIDF is consistently better than BM25, across all four groups of users, using all personalized query expansion methods and in all evaluation metrics. This shows that in terms of frequency-based techniques, more complex

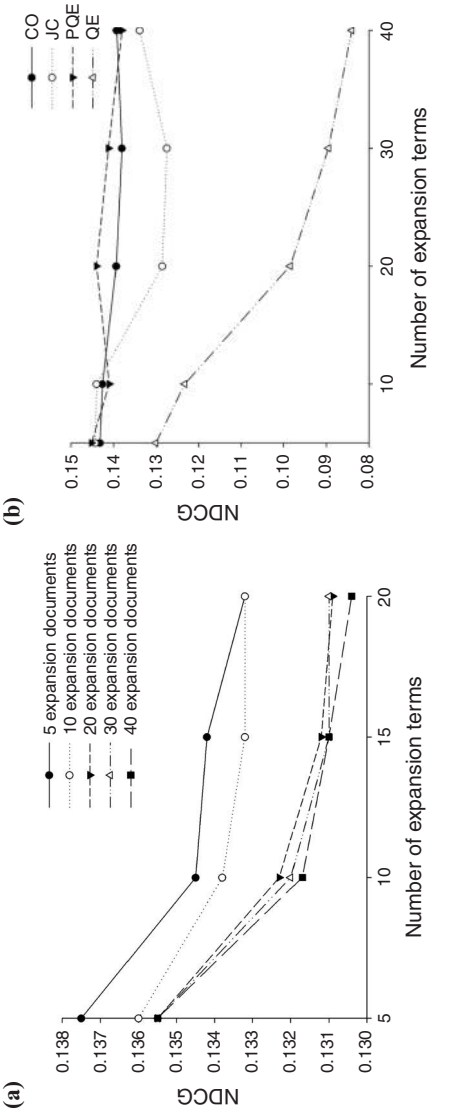

Figure 3.
Parameter settings

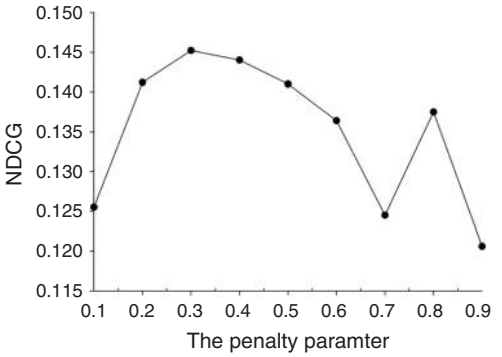

Figure 4.
The impact of varying parameter δ_1

techniques may not work well in representing user profiles. A possible explanation for this result is that complex techniques are tuned using a much larger corpus rather than the small group of documents inside the user profile. It also confirms that simple methods can yield very good results and are fast to compute.

7.1.2 Performance of latent semantic techniques. By examining the results in Table II, several conclusions can be derived. In general, latent semantic techniques work better than frequency-based techniques, with statistically significant results often observed. The best results for every retrieval strategy in every group are produced by the latent semantic techniques. This demonstrates that using semantic-based techniques can find hidden information associated within the text, and in turn produce better user profile representation.

In terms of the latent semantic models themselves, the LDA-1 model works better than the LDA-2 model. It seems that integrating document-topic distributions with topic-word distributions can provide more productive information about the importance of a term inside the user profile. As expected, the bilingual LDA model outperforms the monolingual LDA model in many cases. The only exception is for the CO method. This is probably due to the way the method chooses expansion terms. Some highly weighted terms are not chosen due to their non-co-occurrence with the query terms. Regardless of the only negative result, in most cases the bilingual LDA model can produce more accurate user profile representation for use in personalized CLIR. This confirms that user profiles represented by latent semantic models trained in a cross-lingual manner gained better performance than the models trained monolingually.

7.1.3 Discussion. From the results above, a natural question could be asked: Is it safe to draw the conclusion that using semantic-based techniques is preferable to frequency-based techniques? Although the semantic-based techniques demonstrated better results, they are generally more computationally complex. It is well known that the complexity of the LDA and bilingual models will grow linearly with the number of topics and the number of documents. One possible solution would be only applying them to a document set significantly smaller than the whole corpus, which could contain millions of documents, as has been done here. When the size of the user profile increases, or in other words the user has historical usage information larger than a single LDA can handle, some form of parallelism would be needed.

Second, it is well known that latent semantic-based methods suffer from an incremental build problem. Normally adding new documents to the corpus needs to

Table II.
Main experimental results

	TFIDF			BM25			LDA-1			LDA-2			BLDA		
	P@1	NDCG	MRR	P@1	NDCG	MRR	P@1	NDCG	MRR	P@1	NDCG	MRR	P@1	NDCG	MRR
WIKIgt500															
NOP	0.0614	0.1081	0.0864	0.0614	0.1081	0.0864	0.0614	0.1081	0.0864	0.0614	0.1081	0.0864	0.0614	0.1081	0.0864
JC	*0.0712*	*0.1210*	*0.095*	*0.0712*	*0.1213*	*0.095*	*0.0712*	*0.1235*	*0.0955*	*0.0712*	*0.1230*	*0.0955*	*0.0721*	*0.1246*	*0.0964*
CO	0.0623	*0.1153*	*0.0882*	0.0623	*0.1143*	*0.0869*	0.0632	*0.1169*	*0.0885*	0.0623	*0.1142*	*0.0869*	0.0632	*0.1171*	*0.0893*
PQE	*0.0757*	*0.1224*	*0.1028*	*0.0730*	*0.1213*	*0.1006*	*0.0784*	*0.1275*	*0.1055*	*0.0775*	*0.1258*	*0.1045*	*0.0784*	*0.1297*	*0.1064*
QE	0.0508	0.0851	0.0637	0.049	0.0837	0.0636	0.0518	0.0945	0.0705	0.0508	0.0864	0.0646	0.0523	0.0961	0.0720
WIKI500															
NOP	0.1038	0.1745	0.1411	0.1038	0.1745	0.1411	0.1038	0.1745	0.1411	0.1038	0.1745	0.1411	0.1038	0.1745	0.1411
JC	*0.1138*	*0.1854*	*0.1499*	*0.1138*	*0.1846*	*0.1492*	*0.1138*	*0.1894*	*0.1504*	*0.1118*	*0.1885*	*0.1482*	*0.1138*	*0.1901*	*0.1508*
CO	0.1078	*0.1876*	0.1441	0.1078	*0.1846*	0.1441	0.1118	*0.1884*	0.1455	0.1118	*0.1883*	0.1452	0.1158	*0.1891*	*0.1491*
PQE	*0.1198*	*0.1936*	*0.1568*	*0.1198*	*0.1850*	*0.1555*	*0.1238*	*0.1967*	*0.1571*	*0.1198*	*0.1957*	*0.1570*	*0.1238*	*0.1989*	*0.1594*
QE	0.0439	0.1048	0.0725	0.0439	0.1044	0.0723	0.0599	0.1218	0.0885	0.0599	0.1183	0.0860	0.0619	0.1230	0.0897
WIKI100															
NOP	0.1818	0.2884	0.2421	0.1818	0.2884	0.2421	0.1818	0.2884	0.2421	0.1818	0.2884	0.2421	0.1818	0.2884	0.2421
JC	0.1909	0.2895	0.2440	0.1909	0.2895	0.2440	0.1909	0.2919	*0.2477*	0.1909	0.2918	*0.2475*	*0.2000*	*0.302*	*0.2566*
CO	0.1909	0.2885	0.2426	0.1909	0.2885	0.2426	0.1909	0.2895	0.2446	0.1909	0.2888	0.2434	0.1909	*0.2885*	*0.2426*
PQE	0.1909	*0.2945*	*0.2534*	0.1909	*0.2945*	*0.2534*	0.1919	*0.3045*	*0.2552*	0.1909	*0.3042*	*0.2545*	*0.2009*	*0.3080*	*0.2584*
QE	0.1182	0.2211	0.1745	0.1091	0.2203	0.1712	0.1182	0.2241	0.1763	0.1180	0.2212	0.1717	0.1202	0.2246	0.1766
WIKI150															
NOP	0.2035	0.3150	0.2710	0.2035	0.3150	0.2710	0.2035	0.3150	0.2710	0.2035	0.3150	0.2710	0.2035	0.3150	0.2710
JC	0.2041	*0.3200*	*0.2810*	0.2041	0.3180	*0.2800*	0.2041	*0.3230*	*0.2810*	0.2041	*0.3213*	*0.2800*	0.2041	*0.3252*	*0.2858*
CO	0.2041	0.3180	0.2716	0.2041	0.3180	0.2716	0.2041	*0.3203*	0.2725	0.2041	*0.3185*	0.2716	0.2041	*0.3155*	0.2707
PQE	0.2045	*0.3218*	*0.2830*	0.2042	*0.3200*	*0.2800*	0.2048	*0.3252*	*0.2850*	0.2045	*0.3218*	*0.2830*	*0.2055*	*0.3280*	*0.2883*
QE	0.1224	0.2116	0.1641	0.1224	0.2000	0.1602	0.1224	0.2206	0.1802	0.1224	0.2094	0.1759	0.1429	0.2214	0.1866

Note: Statistically significant results with respect to the NOP approach are marked in italics

"be folded in" to the latent representation. Such incremental addition fails to capture the co-occurrences of the newly added documents (and even ignores all new terms they contain). As such, the quality of the representation will degrade as more documents are added and will eventually require a re-computation of the model. This problem should be able to be avoided by carefully balancing the user's short- and long-term interests. Long-term interests are persistent interests that can be exhibited in the user's search history on the long run, while short-term interests are ephemeral interests that are usually satisfied by a few ad-hoc searches in a relatively shorter period of time. Latent semantic models are most suitable for long-term interests as the content in a user profile is relatively stable. However, at the time of computation and re-computation, such models will perfectly reflect a user's short-term interests.

It should be noted that frequency-based techniques, despite having low computational cost, might lose some important information by only counting the frequencies of terms. They also still suffer from the long- and short-term problems. Careful consideration should be given when choosing which type of user profile representation to use for personalized CLIR.

7.2 Comparison of personalized query expansion techniques

This set of experimental results describes the performance of the techniques used to personalize search, which are shown in Figure 5 for the WIKIgt500 group (as similar results obtained for other groups). As illustrated by the results, the PRF model performed consistently poorly for all evaluation metrics. 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 the large evaluation campaigns. This further demonstrates that using monolingual user profiles 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 23.1 percent 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 (Teevan *et al.*, 2008). The expanded queries can tend to retrieve documents closer to the user profile than the original query. However, different queries have different requirements when it comes to personalization. For some "ambiguous" queries, different users want different results, while for other queries, the majority of users are actually looking for the same result. 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.

The performance between the CO and the JC methods is quite similar. Both methods place more emphasis on the original queries by selecting terms with high co-occurrence statistics with the query terms rather than highly weighted user profile terms. However, they are all beaten by the PQE method. This further confirms that proper use of the profile terms is important in designing any personalized system. Another exciting observation is that in many cases, the personalized query expansion methods, even though tuned for NDCG, can outperform the baselines for all the evaluation metrics, with statistically significant improvements observed frequently.

7.3 Comparison between groups

Next, the performance differences between different groups of users are considered. Figure 6 shows a comparison of performance for the four groups in the context of user

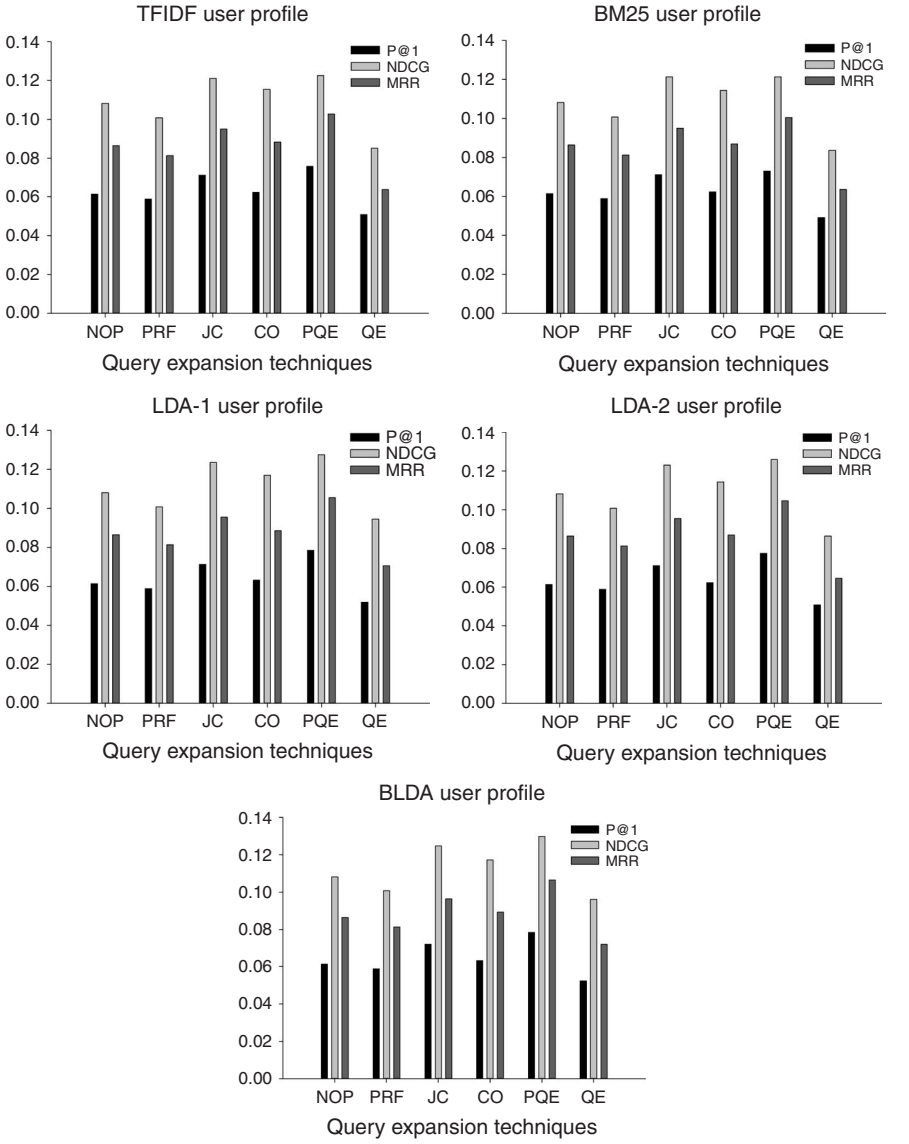

Figure 5.
Comparison of query expansion techniques for the WIKIgt500 group

activities with the NDCG metric (as similar behavior is observed in other metrics). As can be seen, when there is less sufficient data in the user profile, the effectiveness of the personalized approaches will decrease (e.g. see the performance of the PQE method in the WIKI50 group). For less active users, an individual user profile may not be the most suitable way to personalized search. An aggregate user profile could be used in this case. Moreover, latent semantic-based models generally need sufficient data to tell a meaningful story. The cold-start problem, however, can be addressed by matching the user profile to some external resources, such as a reference corpus.

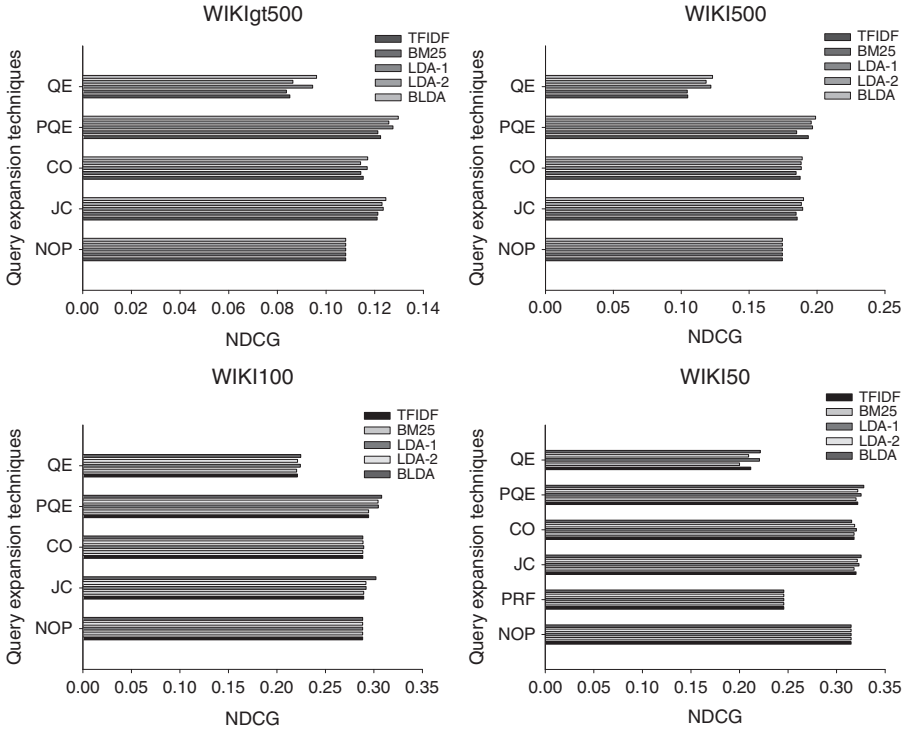

Figure 6. Comparison between groups

7.4 Discussion of the evaluation methodology

In the beginning of the paper we have outlined the difficult problem of evaluating personalized CLIR approaches. The evaluation step is crucial in the development and improvement of systems based upon these approaches. At the same time, the test collections provided by popular large evaluation campaigns do not currently fulfill the requirements of personalized evaluation. The most obvious shortfall is that those collections do not provide individual relevance judgments for each test query. In this research we extended an automatic evaluation methodology previously proposed in the monolingual environment for evaluating personalized CLIR systems. We argue that the evaluation approach can produce repeatable, comparable and generalizable test collections for future research. The documents used in our proposed method can be easily downloaded through Wikipedia dumps[11] or through the crowd-sourced community like DBpedia[12]. The second component of the evaluation framework is the user profiles. The generation of user profiles is achieved through clustering. The technique for document clustering is quite mature and easily adopted. Furthermore, we do not use in-house user queries. In order to maximize the reproducibility of the method, we also employed an automatic query generation method to avoid any bias resulting from human-created queries. Relevance judgments are also straightforward. All in all, just like large evaluation campaigns like CLEF and NTCIR, anyone with moderate knowledge in IR can easily recreate this method and evaluate their own personalization approaches.

With respect to the performance of the different techniques evaluated on the proposed test collection, Vicente-López *et al.* (2015) have stated that the automatic method "[…] is

able to robustly evaluate any given personalization technique, independently of the used retrieval model […]" and "[…] it seems that when the differences in performance between different user profile configurations of a given personalization method are important […]", the automatic method "[…] is able to discriminate among them […]." We also confirmed in our experiments that the differences between different techniques used are clearly illustrated and can be used for a reference to develop more advanced user profile representations and personalization techniques.

As stated earlier, the main goal of the research presented in this paper is to select the best user profile representation techniques and the best personalization strategies for personalized CLIR. So in this paper, the main purpose is not to re-produce the real users' activities by using the simulated users, but rather to compare different personalization performance. As pointed out by Vicente-López *et al.* in their ASPIRE framework, the results obtained using the framework are very close to the real NDCG values obtained from the user study. It is able to discriminate between both different personalization techniques and differences in performance between the different user profile configurations of a given personalization method. In the future, we will use search logs to investigate the search performance between the simulated users and real users.

8. Conclusions and future work

In this paper, a comprehensive study of user profile representation for personalized CLIR has been presented. Techniques based on both frequency and latent semantics are proposed and systematically evaluated. Latent semantic-based techniques demonstrated higher performance than frequency-based techniques, but they still suffer from a high computational cost problem. Careful consideration should be made to balance between effectiveness and efficiency when choosing which type of user profile representation to use. In addition to user profile construction, various query expansion methods are also introduced and compared. Experiments showed that user profiles generated from one language can be used to enhance personalized CLIR, and in general personalized approaches work better than non-personalized approaches. A novel personalized CLIR evaluation methodology is also developed to help lighten the common high barrier in personalized CLIR evaluation. The method aims to ensure repeatable and controlled experiments between different personalized strategies, ensuring comparable measures and generalizable conclusions about them.

We can now comfortably answer the three research questions we proposed in the beginning of the paper:

RQ1. Can we use the profile representation in one language to enhance cross-language search?

Yes, as we can see from the experimental results, user profiles built by either frequency-based methods or latent semantic-based methods in one language can produce better results than non-personalized baselines when suitable personalization strategies are used:

RQ2. Should we take the multilingual dimension of search into consideration when generating user profile representations?

Clearly yes, we have demonstrated in the experimental Section 7.1.2, in most of the cases the bilingual LDA model can produce more accurate user profile representation than monolingual LDA for use in personalized CLIR. This confirms that user profiles

represented by latent semantic models trained in a cross-lingual manner gained better performance than the models trained monolingually:

RQ3. Which user profile representation techniques are most suitable for personalized CLIR?

This question should be viewed objectively. Although the semantic-based techniques for user profile representation demonstrated better results than frequency-based techniques, they are generally more computationally complex. In the contrast, frequency-based techniques have low computational cost, but may lose some important information by only counting the frequencies of terms.

This research continues along several dimensions. Future work is currently being planned to integrate frequency-based and latent semantic-based techniques for representing user profiles. To avoid the query-drift problem, a novel query expansion method will be designed to consider both the terms in the user profile and in the original query. In the current paper, only one type of personalization strategy (query adaptation) has been investigated. The use of results adaptation and a combination of both approaches will be examined in future research. It should be noted that further research is being conducted to further reinforce and establish support for the results described in this paper.

Acknowledgments

The work described in this paper was supported by the National Natural Science Foundation of China under Project No. 61300129, No. 61572187 and No. 61272063, and a project sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, China under Grant Number [2013] 1792. This work is also supported by the ADAPT Centre for Digital Content Technology, which is funded under the Science Foundation Ireland Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund. The authors would like to thank the anonymous reviewers who significantly improved the quality of this manuscript during preparation.

Notes

1. According to http://en.wikipedia.org/wiki/Languages_used_on_the_Internet, nearly half of content provided on the current web is in languages other than English. For example, the estimated percentages of the top ten million websites using various content languages are Russian (6 percent), German (6 percent), Japanese (5 percent), Spanish (4.6 percent), French (4 percent) and Chinese (3.3 percent) and many more. The internet users are multilingual as well. For example, the top three number of internet users are English (28.6 percent), Chinese (23.2 percent) and Spanish (7.9 percent). Note that there are many other facts about the language distributions, interested users could refer to the respected Wikipedia pages and references.
2. www.clef-initiative.eu/
3. http://research.nii.ac.jp/ntcir/index-en.html
4. www.wikipedia.org/
5. www.dmoz.org/
6. www.bing.com/translator
7. http://git.oschina.net/zhzhenqin/paoding-analysis

8. www.bing.com/translator/
9. www.terrier.org
10. This method is included in the Terrier distribution.
11. http://dumps.wikimedia.org/
12. http://wiki.dbpedia.org/

References

Agichtein, E., Brill, E. and Dumais, S. (2006), "Improving web search ranking by incorporating user behavior information", in Efthimiadis, E.N., Dumais, S.T., Hawking, D. and Järvelin, K. (Eds), *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA*, ACM, New York, NY, pp. 19-26.

Amati, G. and Rijsbergen, C.J.V. (2002), "Probabilistic models of information retrieval based on measuring the divergence from randomness", *ACM Transactions on Information Systems*, Vol. 20 No. 4, pp. 357-389.

Ambati, V. and Rohini, U. (2006), "Using monolingual clickthrough data to build cross-lingual search systems", in Gey, F.C., Kando, N., Lin, C.-Y. and Peters, C. (Eds), *New Directions in Multilingual Information Access Workshop of SIGIR 2006, Seattle, WA*, ACM, New York, NY, pp. 28-35.

Azzopardi, L., de Rijke, M. and Balog, K. (2007), "Building simulated queries for known-item topics: an analysis using six European languages", in Kraaij, W., de Vries, A.P., Clarke, C.L.A., Fuhr, N. and Kando, N. (Eds), *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam*, ACM, New York, NY, pp. 455-462.

Baeza-Yates, R. and Ribeiro-Neto, B. (2011), *Modern Information Retrieval: The Concepts and Technology Behind Search*, 2nd ed., Addison-Wesley Professional, NewYork, NY.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent Dirichlet allocation", *Journal of Machine Learning Research*, Vol. 3 No. 1, pp. 993-1022.

Boyd-Graber, J. and Blei, D.M. (2009), "Multilingual topic models for unaligned text", *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, *AUAI Press, Montreal, Corvallis, OR*, pp. 75-82.

Brusilovsky, P. and Millán, E. (2007), "User models for adaptive hypermedia and adaptive educational systems", in Brusilovsky, P., Kobsa, A. and Nejdl, W. (Eds), *The Adaptive Web*, Springer Berlin Heidelberg, Berlin, pp. 3-53.

Cai, F., Liang, S. and de Rijke, M. (2014), "Personalized document re-ranking based on Bayesian probabilistic matrix factorization", in Geva, S., Trotman, A., Bruza, P., Clarke, C.L.A. and Järvelin, K. (Eds), *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast*, ACM, New York, NY, pp. 835-838.

Cao, G., Gao, J., Nie, J.-Y. and Bai, J. (2007), "Extending query translation to cross-language query expansion with Markov chain models", *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, Lisbon*, ACM, New York, NY, pp. 351-360.

Cao, G., Nie, J.-Y., Gao, J. and Robertson, S. (2008), "Selecting good expansion terms for pseudo-relevance feedback", in Myaeng, S.-H., Oard, D.W., Sebastiani, F., Chua, T.-S. and Leong, M.-K. (Eds), *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore*, ACM, New York, NY, pp. 243-250.

Carpineto, C. and Romano, G. (2012), "A survey of automatic query expansion in information retrieval", *ACM Computing Surveys*, Vol. 44 No. 1, pp. 1-50.

Chirita, P.-A., Firan, C.S. and Nejdl, W. (2007), "Personalized query expansion for the web", in Kraaij, W., de Vries, A.P., Clarke, C.L.A., Fuhr, N. and Kando, N. (Eds), *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam*, ACM, New York, NY, pp. 7-14.

Cimiano, P., Schultz, A., Sizov, S., Sorg, P. and Staab, S. (2009), "Explicit versus latent concept models for cross-language information retrieval", in Boutilier, C. (Ed.), *Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, CA*, AAAI Press, Palo Alto, CA, pp. 1513-1518.

Efthimiadis, E.N. (1995), "User choices: a new yardstick for the evaluation of ranking algorithms for interactive query expansion", *Information Processing & Management*, Vol. 31 No. 4, pp. 605-620.

Egozi, O., Markovitch, S. and Gabrilovich, E. (2011), "Concept-based information retrieval using explicit semantic analysis", *ACM Transactions on Information Systems*, Vol. 29 No. 2, pp. 1-34.

Gao, J., Nie, J.-Y., Xun, E., Zhang, J., Zhou, M. and Huang, C. (2001), "Improving query translation for cross-language information retrieval using statistical models", in Croft, W.B., Harper, D.J., Kraft, D.H. and Zobel, J. (Eds), *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, LA*, ACM, New York, NY, pp. 96-104.

Ghorab, M.R., Zhou, D., Lawless, S. and Wade, V. (2012), "Multilingual user modeling for personalized re-ranking of multilingual web search results", in Herder, E., Yacef, K., Chen, L. and Weibelzahl, S. (Eds), *CEUR Workshop Proceeding of UMAP 2012, Montreal*, Springer, Berlin, pp. 1-4.

Ghorab, M.R., Zhou, D., O'Connor, A. and Wade, V. (2013), "Personalised information retrieval: survey and classification", *User Modeling and User-Adapted Interaction*, Vol. 23 No. 4, pp. 381-443.

Ghorab, M.R., Zhou, D., Steichen, B. and Wade, V. (2011), "Towards multilingual user models for personalized multilingual information retrieval", in Agosti, M., De Luca, E.W., Lawless, S. and Leveling, J. (Eds), *Proceedings of the First Workshop on Personalised Multilingual Hypertext Retrieval, Eindhoven*, ACM, New York, NY, pp. 42-49.

Gollins, T. and Sanderson, M. (2001), "Improving cross language retrieval with triangulated translation", in Croft, W.B., Harper, D.J., Kraft, D.H. and Zobel, J. (Eds), *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, LA*, ACM, New York, NY, pp. 90-95.

Hsu, C.C., Li, Y.T., Chen, Y.W. and Wu, S.H. (2008), "Query expansion via link analysis of Wikipedia for CLIR", *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NII, Tokyo*, pp. 125-131.

Järvelin, K. and Kekäläinen, J. (2002), "Cumulated gain-based evaluation of IR techniques", *ACM Transactions on Information Systems*, Vol. 20 No. 4, pp. 422-446.

Lin, M.-C., Li, M.-X., Hsu, C.-C. and Wu, S.-H. (2010), "Query Expansion from Wikipedia and Topic Web Crawler on CLIR", *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NII, Tokyo*, pp. 101-106.

Littman, M., Dumais, S. and Landauer, T. (1998), "Automatic cross-language information retrieval using latent semantic indexing", in Grefenstette, G. (Ed.), *Cross-Language Information Retrieval*, Springer, New York, NY, pp. 51-62.

Liu, Y., Niculescu-Mizil, A. and Gryc, W. (2009), "Topic-link LDA: joint models of topic and author community", in Danyluk, A.P., Bottou, L. and Littman, M.L. (Eds), *Proceedings of the 26th Annual International Conference on Machine Learning, Montreal*, ACM, New York, NY, pp. 665-672.

McNamee, P. and Mayfield, J. (2002), "Comparing cross-language query expansion techniques by degrading translation resources", in Beaulieu, M., Baeza-Yates, R. and Myaeng, S.H. (Eds), *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere*, ACM, New York, NY, pp. 159-166.

Micarelli, A. and Sciarrone, F. (2004), "Anatomy and empirical evaluation of an adaptive web-based information filtering system", *User Modeling and User-Adapted Interaction*, Vol. 14 Nos 2-3, pp. 159-200.

Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A. and McCallum, A. (2009), "Polylingual topic models", in Koehn, P. and Mihalcea, R. (Eds), *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore*, ACL, New York, NY, pp. 880-889.

Ni, X., Sun, J.-T., Hu, J. and Chen, Z. (2011), "Cross lingual text classification by mining multilingual topics from Wikipedia", in King, I., Nejdl, W. and Li, H. (Eds), *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, Hong Kong*, ACM, New York, NY, pp. 375-384.

Nie, J.-Y. (2010), *Cross-Language Information Retrieval*, Morgan and Claypool Publishers, San Rafael, CA.

Oard, D. (1998), "A comparative study of query and document translation for cross-language information retrieval", in Farwell, D., Gerber, L. and Hovy, E. (Eds), *Machine Translation and the Information Soup*, Springer Berlin Heidelberg, Berlin, pp. 472-483.

Ruiz, M.E. and Chin, P. (2010), "Users' image seeking behavior in a multilingual tag environment", in Peters, C., Caputo, B., Gonzalo, J., Jones, G.J.F., Kalpathy-Cramer, J., Müller, H. and Tsikrika, T. (Eds), *Multilingual Information Access Evaluation II. Multimedia Experiments*, Springer Berlin Heidelberg, Berlin, pp. 37-44.

Ruthven, I. (2003), "Re-examining the potential effectiveness of interactive query expansion", in Callan, J., Hawking, D. and Smeaton, A. (Eds), *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto*, ACM, New York, NY, pp. 213-220.

Shen, X., Tan, B. and Zhai, C. (2005), "Implicit user modeling for personalized search", in Herzog, O., Schek, H.-J., Fuhr, N., Chowdhury, A. and Teiken, W. (Eds), *Proceedings of the 14th ACM International Conference on Information and Knowledge Management, Bremen*, ACM, New York, NY, pp. 824-831.

Sorg, P. and Cimiano, P. (2008), "Cross-lingual information retrieval with explicit semantic analysis", working notes for the CLEF 2008 Workshop, Aarhus.

Steichen, B., Ghorab, M.R., O'Connor, A., Lawless, S. and Wade, V. (2014), "Towards personalized multilingual information access – exploring the browsing and search behavior of multilingual users", in Dimitrova, V., Kuflik, T., Chin, D., Ricci, F., Dolog, P. and Houben, G-J. (Eds), *Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization, Aalborg*, Springer, Berlin, pp. 435-446.

Teevan, J., Dumais, S.T. and Horvitz, E. (2005), "Personalizing search via automated analysis of interests and activities", in Baeza-Yates, R.A., Ziviani, N., Marchionini, G., Moffat, A. and Tait, J. (Eds), *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador*, ACM, New York, NY, pp. 449-456.

Teevan, J., Dumais, S.T. and Liebling, D.J. (2008), "To personalize or not to personalize: modeling queries with variation in user intent", in Myaeng, S.-H., Oard, D.W., Sebastiani, F., Chua, T.-S. and Leong, M.-K. (Eds), *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore*, ACM, New York, NY, pp. 163-170.

Vassilakaki, E., Johnson, F., Hartley, R.J. and Randall, D. (2009), "Users' perceptions of searching in flickling", working notes for the CLEF 2009 Workshop, Corfu.

Vicente-López, E., de Campos, L., Fernández-Luna, J., Huete, J., Tagua-Jiménez, A. and Tur-Vigil, C. (2015), "An automatic methodology to evaluate personalized information retrieval systems", *User Modeling and User-Adapted Interaction*, Vol. 25 No. 1, pp. 1-37.

Vulić, I., De Smet, W. and Moens, M.-F. (2013), "Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora", *Information Retrieval*, Vol. 16 No. 3, pp. 331-368.

Wang, C. and Blei, D.M. (2011), "Collaborative topic modeling for recommending scientific articles", in Apté, C., Ghosh, J. and Smyth, P. (Eds), *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA*, ACM, New York, NY, pp. 448-456.

Wei, X. and Croft, W.B. (2006), "LDA-based document models for ad-hoc retrieval", in Efthimiadis, E.N., Dumais, S.T., Hawking, D. and Järvelin, K. (Eds), *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA*, ACM, New York, NY, pp. 178-185.

White, R.W. and Marchionini, G. (2007), "Examining the effectiveness of real-time query expansion", *Information Processing & Management*, Vol. 43 No. 3, pp. 685-704.

White, R.W., Chu, W., Hassan, A., He, X., Song, Y. and Wang, H. (2013), "Enhancing personalized search by mining and modeling task behavior", in Schwabe, D., Almeida, V.A.F., Glaser, H., Baeza-Yates, R.A. and Moon, S.B. (Eds), *Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro*, ACM, New York, NY, pp. 1411-1420.

Xu, S., Bao, S., Fei, B., Su, Z. and Yu, Y. (2008), "Exploring folksonomy for personalized search", in Myaeng, S.-H., Oard, D.W., Sebastiani, F., Chua, T.-S. and Leong, M.-K. (Eds), *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore*, ACM, New York, NY, pp. 155-162.

Zhao, B. and Xing, E.P. (2006), "BiTAM: bilingual topic AdMixture models for word alignment", in Calzolari, N., Cardie, C. and Isabelle, P. (Eds), *Proceedings of the COLING/ACL on Main Conference Poster Sessions, Sydney*, ACL, New York, NY, pp. 969-976.

Zhou, D., Lawless, S. and Wade, V. (2012a), "Improving search via personalized query expansion using social media", *Information Retrieval*, Vol. 15 Nos 3-4, pp. 218-242.

Zhou, D., Lawless, S., Min, J. and Wade, V. (2010), "A late fusion approach to cross-lingual document re-ranking", in Huang, J., Koudas, N., Jones, G.J.F., Wu, X., Collins-Thompson, K. and An, A. (Eds), *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto*, ACM, New York, NY, pp. 1433-1436.

Zhou, D., Truran, M., Brailsford, T., Wade, V. and Ashman, H. (2012b), "Translation techniques in cross-language information retrieval", *ACM Compututing Surveys*, Vol. 45 No. 1, pp. 1-44.

Zhou, D., Truran, M., Liu, J., Li, W. and Jones, G. (2014), "Iterative refinement methods for enhanced information retrieval", *International Journal of Intelligent Systems*, Vol. 29 No. 4, pp. 341-364.

Corresponding author
Dong Zhou can be contacted at: dongzhou1979@hotmail.com