Information Sciences 430–431 (2018) 200–215

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

An iterative method for personalized results adaptation in cross-language search

Dong Zhou[a,*], Wenyu Zhao[a], Xuan Wu[a], Séamus Lawless[b], Jianxun Liu[a]

[a] School of Computer Science and Engineering & Key Laboratory of Knowledge Processing and Networked Manufacturing, Hunan University of Science and Technology, Xiangtan, Hunan 411201, China
[b] ADAPT Centre, Knowledge and Date Engineering Group, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland

ARTICLE INFO

Article history:
Received 16 September 2016
Revised 13 October 2017
Accepted 22 November 2017
Available online 22 November 2017

Keywords:
Personalized cross-language information retrieval
Results adaptation
User model
Iterative method

ABSTRACT

On today's Web, people often desire to not only retrieve results which are of relevance to their query, but for those results to be of particular relevance to them as an individual. In most personalized search systems, the scores obtained from different rankers are linearly combined to provide the personalized ranked list. Moreover, when compared to the personalization research in monolingual web search, relatively few studies extend to the cross-language domain. In this paper we investigate the personalized results adaptation problem in the context of cross-language web search. The main contribution of this research is a novel iterative ranking method based on document associations obtained from an initial ranker. The method assumes that results retrieved by non-personalized rankers and personalized rankers mutually reinforce each other, rather than being used in linear combination. The method is applied in a personalized cross-language search scenario on a semi-automatically constructed test collection and a real-world dataset. The experimental results suggest that the proposed personalized result adaptation method can produce better results than previous approaches for cross-language web search. The results also prove that the semi-automatically constructed test collection can be used as an alternative dataset for evaluation in the absence of available real-world datasets.

1. Introduction

In recent years personalized search systems have been extensively studied in the literature [30,32]. Such systems do not retrieve results that are solely relevant to the query, but rather, results that are also relevant to the individual user's interests, preferences, needs etc. These systems generally pass through three main steps in order to provide their personalized service [17]: (1) information gathering, using tools to collect information about users and their usage history; (2) information representation of user context, which is usually stored as the user model [33]; and (3) personalization implementation, which usually takes the form of different approaches to adapt a user's query or results [12,16]. All of these steps try to minimize user effort and maximize user satisfaction.

The results adaptation approach is commonly used for personalization in many monolingual search systems. Adaptation of result lists can be performed by result scoring, result re-ranking or result filtering. There are three groups of techniques

[*] Corresponding author.
E-mail address: dongzhou@hnust.edu.cn (D. Zhou).

https://doi.org/10.1016/j.ins.2017.11.044

where the adaptation factors related to particular users can be obtained. The first group of techniques uses lexical matching between the documents and terms stored in the user model [33]. In this way, the selected terms from the user model can be viewed as the user's perspective of the query, in contrast to the initial ambiguous query that was issued by the user. However, lexical matching may ignore the semantic information, so researchers have attempted to map the documents and user models onto external knowledge bases. Among this second group of techniques, ODP categories or WordNet [13,28] and folksonomies [6,40,42–44,47] are the most popular choices of external knowledge. Semantic information can also be obtained by using latent semantic models. This third group of techniques use models, such as matrix factorization and Latent Dirichlet Allocation (LDA) [4], to calculate the latent topics and then map the user models and documents onto these topics to perform personalization [5,9].

In the literature [5,7,9,10,40,42–44], the adaptive factors that are used to score the documents according to the users' needs are linearly combined with information retrieval factors (i.e. the query term matching scores) in the scoring function. This combination is achieved using additional parameters to control the degree of personalization. However, the linearly combined scores may miss information when used to perform results adaptation. First, the associations between documents are not considered when personalizing the search results. The documents related to an initial query are not isolated; they are connected and can provide valuable information when calculating the adaptation scores. Second, the relationships between the scores acquired by non-personalized rankers and personalized rankers are not fully explored. Obviously, one set of scores may influence the other and vice versa. In addition, if one decides to add more adaptation scores into the final ranking scores, it is inevitable that the combination parameters will also grow linearly. Tuning such parameters is not easy [1,24,39]. Finally, when compared to the personalization research in monolingual web search, relatively fewer studies extend to the cross-language domain. There is only a single study using the lexical matching approach to calculate adaptation scores for cross-language search [16]. Clearly further investigations are required.

In this paper we address the personalized search problem through result adaptation techniques. We propose a novel iterative method based on document associations obtained from an initial ranker. The intuition behind this method is: *similar documents are likely to have similar scores with respect to both non-personalized and personalized factors*. The method assumes that results retrieved by non-personalized rankers and personalized rankers mutually reinforce each other rather than being used in linear combination. Updating one set of scores will iteratively propagate to other sets of scores via pairwise document relationships. To exploit these relationships, we adjust multiple scores using a function which regularizes the smoothness of document associations over a connected graph. The method also goes beyond bag-of-words based models (i.e. lexical matching). It considers latent topics derived from documents, queries and user models. In our method, we can simultaneously process multiple adaptation scores without complex parameter tuning procedures.

In order to evaluate the proposed method, we apply it in the scenario of personalized cross-language search by using two different test collections. One collection is semi-automatically constructed, motivated by an automatic methodology to evaluate personalized search systems [38] and a method to conduct known-item searches [2]. Another collection is constructed from a real click-through log. The motivation for the experimental procedure is that there are no readily available test collections for cross-language search which contain user-oriented information. Empirical evaluation performed with the two test collections attests to the effectiveness of our method. Specifically, retrieval performance is substantially better than that attained by using the linear combination of different adaptation features for cross-language personalization. Furthermore, the experimental results also prove that the semi-automatically constructed test collection can be used as an alternative basis for evaluation in the absence of real world datasets.

The contributions of this paper can be summarized as follows:

i. *We tackle the challenge of personalized results adaptation in a novel way by assuming results retrieved by non-personalized rankers and personalized rankers mutually reinforce each other rather than being linearly combined.*
ii. *We propose a novel iterative method based on document associations obtained from an initial ranker that can simultaneously process multiple adaptation scores.*
iii. *We apply the proposed method in the context of personalized cross-language search by suggesting an evaluation methodology to help lower the high barrier to the evaluation of such systems. The approach 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 is summarized in Section 2. Section 3 presents details of the proposed personalized results adaptation technique. Section 4 demonstrates the procedures used in the calculation of adaptation scores. In Section 5, a report on a series of experiments performed to evaluate the personalization strategies is provided. Also the procedure for building test collections for personalized cross-language search and other experimental settings are outlined. Finally, Section 6 concludes the paper and outlines areas of future research.

2. Related work

There is a significant volume of published research in the area of monolingual personalized search systems [30,32]. One common approach is result adaptation. Adaptation of result lists can be performed by result scoring, result re-ranking, or result filtering. Result re-ranking takes place after an initial set of documents have been retrieved by the system, where an additional ranking round is performed to re-order documents based on certain adaptation aspects (e.g. displaying certain documents at higher ranks in the result list based on the user's interests) [9,16,33]. Result filtering can be considered as a special case (or a further step) of result re-ranking, where after the result list is sorted in descending order of relevance scores, results that fall below a certain threshold are not displayed to the user [26]. Result scoring involves incorporating adaptation features directly into the primary scoring function of the retrieval component of the system [42,43]. It is worth noting that, the distinction between the three approaches is not firm, they are often used together.

A commonly used approach [5,7,9,10,40,42–44] is for the adaptive factors that are used to score the documents according to the users' needs to be linearly combined together with the information retrieval factors in the scoring function as follows:

$$S_{final}(d) = (1 - \lambda)S_{term}(q, d) + \lambda S_{adapt}(u, d) \tag{1}$$

where the final ranking score $S_{final}(d)$ for a particular document d is a linear combination of the query (q) term matching score $S_{term}(q,\ d)$ and the user (u) adaptation score $S_{adapt}(u,\ d)$, λ is used to control the degree of personalization. However, this method may miss information when performing results adaptation, such as document associations and the relationships between different set of scores. Moreover, as pointed out by previous researchers, although good results are achieved in some situations, the linear combination method has not yet been shown to produce reliable improvement [1,39]. The same problem exists in related methods such as Weighted Borda-fuse [1]. The weights of documents from different rankers are calculated independently and by using heuristics [24], which means that if we use multiple $S_{term}(q,\ d)$ and $S_{adapt}(u,\ d)$ scores, the parameter tuning procedure is difficult, as simultaneously adjusting multiple parameters is complex and prone to errors.

The main novelty between different personalization approaches lies in changing the adaptation score $S_{adapt}(u,\ d)$. There are three groups of techniques where the adaptation scores can be obtained. The first group of techniques uses lexical matching between the documents and terms stored in the user model [33]. In this way, the selected terms from the user model can be viewed as a user's perspective of the query in contrast to the initial ambiguous query issued by the user. For example, Shen et al. [32] re-adapt search results according to a standard vector-space retrieval model and scores obtained based on the similarity of the result and the current user's information-need vector, in which the user's short-term interests are stored lexically.

However, lexical matching may miss or ignore semantic information, and as a result, researchers have attempted to map the documents and user models onto external knowledge bases. Among this second group of techniques, folksonomies [40,42–44,47] and ODP categories [13] are the most popular choices of external knowledge. MiSearch [33] performs personalization by first mapping Google search results to ODP categories, then by comparing concepts inside the user model to re-adapt the results. Similarly, Chirita et al. [13] calculated the distance between a user profile defined using ODP topics and the set of ODP topics covered by each set of search results returned in regular web search. The authors in [37] investigated how the ranking of search engine results can be improved with respect to users if the users' social information is taken into consideration. Folksonomies can also be used as a test bed for personalized results adaptation in addition to enhancing the user models. Claypool et al. [14] first introduced the idea and Xu et al. [44] developed a well-known personalization approach to learn about users' interests from their bookmarks and tags, then re-adapt the results according to the topical relevance of documents and users' interests. Wang and Jin [40] explored gathering data from multiple online social systems for search results adaptation. There are many other external knowledge bases that can be chosen, such as WordNet [28]. Cai et al. [10] model query relevance measurement and user relevance measurement as fuzzy satisfaction problems. Verbal context [42] and sentiment analysis [43] are also utilized.

Semantic information can also be obtained using latent semantic models. This group of techniques uses models, such as matrix factorization, LDA and more recently word embeddings, to calculate the latent topics, and then map the user models and documents onto these topics to perform personalization. Cai et al. [9] used Bayesian probabilistic Matrix Factorization to predict the relevance of a document for a query as well as the preference of a user for a document in order to readapt the web search results. Sun et al. [35] represents click-through data by a 3-order tensor in order to apply the decomposition technique to capture the latent factors that govern the relations among users, queries and Web pages for results adaptation. Bouadjenek et al. [5] used an LDA based method to compute the similarity between document topics and the topics of the user model in their paper. Zhou and Wade [49] also proposed an LDA-based method for non-personalized search, this method can be easily extended to personalized search (See Section 4).

All techniques mentioned above are performed in monolingual personalized search scenarios. Relatively few studies extend these approaches to the cross-language domain. In fact, most studies in cross-language search are non-user focused, although query adaptation and results adaptation have been thoroughly studied [19,47]. Steichen et al. [34] presented a survey of polyglot users to analyze their multilingual proficiency and browsing/search language preferences in personalized multilingual information access. There are few examples of research that investigate multilingual user models to be used in personalized multilingual IR [16–18]. Specifically, the personalized results adaptation problem was studied in [16]. A multilingual user model was used and a lexical-match-based results adaptation was adopted. However, the method and comparison are both rather simplistic. Only a non-personalized baseline was considered. Moreover, due to the difficulties involved in conducting user-based studies, the experiments conducted in those papers are on a much smaller scale. In contrast, the current paper presents a more complex method for personalized results adaptation when compared to several state-of-the-art methods. We also propose an evaluation framework to help lighten the common high barrier in personalized cross-language search evaluation.

Our proposed method is directly related to the famous TextRank method [27]. TextRank is a graph-based ranking model for text processing, it has been successfully utilized in many natural language processing applications such as keyword and sentence extraction [3] and personalization [31], where the user model contains automatically extracted keywords. Our method proposed here has certain differences with respect to the TextRank method. Firstly, TextRank and the successive usage of the method are mainly for text units, such as words and sentences, however, to the authors' knowledge, no usage has yet been reported at the document level. Secondly, TextRank weights the graph according to the PageRank scheme, which is different from our method (we used the normalized Laplacian). More importantly, TextRank only considers one set of scores, while in our case, we simultaneously consider multiple sets of scores, which is the main novelty of our proposed method. In Section 3.3 we detail the connection between our method and the TextRank method.

3. Iterative method for results adaptation

In this section, we describe a novel iterative method based on document associations obtained from an initial non-personalized ranker. The method assumes that results retrieved by non-personalized rankers and personalized rankers mutually reinforce each other rather than being linearly combined. We will show that updating one set of scores will iteratively propagate to other sets of scores via pairwise document relationships. To exploit these relationships, we adjust multiple scores using a function which regularizes the smoothness of document associations over a connected graph. The intuition behind this method is: *similar documents are likely to have similar scores with respect to both non-personalized and personalized factors*. In other words, the multiple scores for a particular user are adjusted by context enhancing and weighting propagation. The neighbors of a particular document in a connected graph will have similar scores as that document. In addition, the final ranking scores are partially restrained by both non-personalized and personalized factors.

3.1. The method

Let $D = \{d_1, d_2, \dots, d_m\}$ denote the set of documents to be searched. Given a query q, a set of initial results $D_{init} \in D$ are returned by an initial ranker ($S_{term}(q,\ d)$). However, this ranker is non-personalized and the final ranking is usually unsatisfactory to the end user. The purpose of our method is to re-order a set of documents D'_{init} so as to improve retrieval accuracy and user satisfaction at the very top ranks of the final results.

Further, Let $G = (V, E)$ be a connected graph, wherein vertex set $V = D'_{init}$, and edges E correspond to the pairwise document relationships between documents. D'_{init} represents a merge of top documents (i.e. D_{init}) returned by different rankers. In this paper, we consider three different rankers, $S_{term}(q,\ d)$,$S_{latent}(q,\ d)$ and $S_{user}(q,\ d)$. The weights on these edges are calculated using the Jensen–Shannon divergence [23] between documents. We also assume an $n \times n$ symmetric weight matrix A on the edges of the graph, so that a_{ij} denotes the weight between documents d_i and d_j and n is the size of D'_{init}. We further define M as a diagonal matrix with entries $M_{ii} = \sum_j a_{ij}$ and construct the matrix $W = M^{-1/2}AM^{-1/2}$. In general, there are no constraints to construct the weight matrix A, which is a symmetric matrix. The normalized graph Laplacian of matrix A is given by W, which is commonly used in graph-based machine learning approaches [50].

Our method can be viewed as estimating a ranking function on G, subject to the following: 1) it is close to the given scores on the top-ranked documents produced by different non-personalized and personalized rankers, and 2) it is smooth on the whole graph. Here we propose an iterative method similar to label propagation [46,50], which can iteratively propagate the scores of the top-ranked documents produced by different rankers to the remaining documents on a constructed graph.

Let $\bar{\mathcal{F}}$ denote the set of ranking functions defined on V, $\forall f \in \bar{\mathcal{F}}$ that can assign a scoref_i to every document d_i. In each propagation step, we let each document absorb a fraction of score information from its neighborhood and retrain some score information of its initial state (at the moment, we only consider the scores produced by the initial ranker). Therefore, the score of document d_i at time $k+1$ becomes:

$$f_i^{k+1} = \alpha \sum_j w_{ij} f_j^k + (1-\alpha) y_i \tag{2}$$

where $0 < \alpha < 1$ is the fraction of score information that d_i receives from its neighbors, $w_{ij} \in W$. Let $\vec{y} = \{y_1, y_2, \dots, y_n\}^T$ with $y_i = S_{term}(q, d_i)$. $\vec{f^k} = \{f_1^k,\ f_2^k, \dots, f_n^k\}$ is the ranking scores vector at iteration k, and $\vec{f^0} = \vec{y}$. Then, we can rewrite our iteration Eq. (2) as:

$$\vec{f}^{k+1} = \alpha W \vec{f^k} + (1-\alpha)\vec{y} \tag{3}$$

We will use (3) to update the scores of the documents until convergence. Here, "convergence" means that the final ranking scores of the document will not change in several successive iterations (i.e. results at iteration $k+1$ has no change with the results at iteration k). Note that we do not use Eq. (3) to compute our final results as we used multiple sets of scores simultaneously. We used an extension of the above method for our iterative method, as described below.

It is easy to change the single set of scores produced by one ranker $\vec{f^0}$, as used above, to multiple sets of scores produced by different rankers. Suppose there are a set of scores, c, by various non-personalized rankers and personalized rankers. Let $\mathcal{M}$ be a set of $n \times c$ matrices with non-negative real-valued entries. Any matrix $F = [F_1^T, F_2^T, \dots, F_n^T]^T \in \mathcal{M}$ corresponds to a specific ranking function on V that scores d_i as $y_i = argmax_{j \le c} F_{ij}$. Initially we set $F_0 = Y$, where, in this paper, $Y = [S_{term}(q, d),\ S_{latent}(q, d), S_{user}(q, d)]$. $S_{latent}(q,\ d)$ denotes the ranking scores calculated by the latent semantic-based ranker, and $S_{user}(q,\ d)$ denotes the ranking scores calculated by the personalized ranker. We will describe each of the rankers in

Table 1
Iterative method for personalized results adaptation.

```
Input: A set of initial results $D'_{init} \in D$,
  The initial score matrix $Y$,
  The constant $\alpha$.
Output: The final ranking scores of documents.
1: Form the affinity matrix $A$, which is defined by Jensen–Shannon divergence between documents retrieved by the initial ranker;
2: Construct the propagation matrix $W = M^{-1/2}AM^{-1/2}$;
3: Iterate $F_{t+1} = \alpha W F_t + (1 - \alpha)Y$ until convergence;
4: Let $F^*$ be the limit of the sequence $F_t$. Output the scores of each document $d_i$ by $y_i = argmax_{j \le c} F^*_{ij}$.
```

detail in Section 4. We keep the scores of the β top-ranked documents for each ranker and the rest of scores are set to zero. The main procedure of our iterative method is summarized in Table 1.

We now prove that the sequence F_t will converge to:

$$F^* = (1 - \alpha)(I - \alpha W)^{-1} Y \tag{4}$$

when $t \to \infty$ and $F_0 = Y$.

Proof. By $F_{t+1} = \alpha W F_t + (1 - \alpha)Y$ and the initial condition that $F_0 = Y$, we have:

$$F_t = (\alpha W)^{t-1} Y + (1 - \alpha) \sum_{i=0}^{t-1} (\alpha W)^i Y \tag{5}$$

Since $0 < \alpha < 1$ and the eigenvalues of W in $[-1, 1]$, so that:

$$\lim_{t \to \infty} (\alpha W)^{t-1} = 0 \text{ and } \lim_{t \to \infty} \sum_{i=0}^{t-1} (\alpha W)^i Y = (I - \alpha W)^{-1} Y$$

Hence the sequence F_t will converge to:

$$F^* = \lim_{t \to \infty} F_t = (1 - \alpha)(I - \alpha W)^{-1} Y$$

which proves the assumption. □

3.2. Computational efficiency

The computational cost of the iterative method mainly consists of two parts, one is for graph construction and the other is for the iterative procedure. We can easily derive that the computational cost of graph construction is $O(|D'_{init}|^2)$, where $|D'_{init}|$ represents the number of documents in D'_{init}. The cost of the iterative procedure is $O(|D'_{init}|^2 \times C \times i)$, where C is the set of scores by various non-personalized rankers and personalized rankers. i is the number of iterations. Obviously the iterative procedure is much more rapid than the graph construction. In practice the method converges very quickly, we demonstrate the performance variation in the iterative process in the experimental section (Section 5.4.3). Given the fact that there are only a few documents in D'_{init} (usually less than 1000 in a proper information retrieval experiment, and even fewer in real web search practice) and the graph only has to be constructed once, the cost is very manageable.

3.3. Connection with TextRank

As mentioned in the related work section, our proposed method has a direct connection to the TextRank [27] method proposed by Mihalcea and Tarau. In this subsection, we formally discuss this connection and show that TextRank produces the same ranked list if we set our method to use only one set of scores. TextRank uses a weighted version of PageRank to rank the text unit. We can re-write the equation from Section 2.2 of the TextRank paper [27] as follows:

$$P = (1 - \beta)U + \beta M'^{-1} W' \tag{6}$$

where U represents the matrix with all entries equal to $1/n$. This can be interpreted as a probability β of transition to an adjacent vertex, and a probability $1 - \beta$ of randomly jumping to any point on the graph uniform. We can easily obtain the stationary distribution P‘ of the random walk used in TextRank:

$$P\text{'} = 1M'/Vol_G \tag{7}$$

It is worth noting that we only consider one set of initial scores here, and we do not consider the actual weights. 1 denotes the $1 \times n$ vector with all entries equal to 1, and Vol_G denotes the volume of G, which is given by the sum of vertex degrees. By obtaining Eq. (7) we consider that $1M'P = 1M'[(1 - \beta)U + \beta M'^{-1} W'] = (1 - \beta)1M'U + \beta 1M'M'^{-1}W' = (1 - \beta)1M' + \beta 1W' = (1 - \beta)1M' + \beta 1M' = 1M'$.

In our iterative method, if we only consider one set of scores and discard the initial scores, our iterative step becomes: $\vec{f}^{k+1} = \alpha W \vec{f}^k$. The sequence $\{\vec{f}^k\}$ converges to the principle eigenvector of W. Now we let 1 denote the $n \times 1$ vector with all entries equal to 1, then we can obtain:

$$\vec{f}^* = M^{1/2}1 \tag{8}$$

by considering $WM^{1/2}1 = M^{-1/2}AM^{-1/2}M^{1/2}1 = M^{-1/2}W1 = M^{-1/2}M1 = M^{1/2}1$, which is the eigenvector of W.

By comparing Eq. (7) and (8), we can see that P' and $\vec{f}^*$ give the same ranked list (see also [15]). This means that TextRank produces the same ranked list if we set our method to use only one set of scores. However, the main novelty in our method is that we simultaneously consider multiple sets of scores and calculate the final ranked list according to the reinforcements of those scores. This means that our method can be regarded as equivalent to choosing the maximum after several TextRank approaches have been simultaneously run. Furthermore, TextRank and the successive usage of the method are mainly for text units, such as words and sentences, however, to the authors' knowledge, no usage has been reported yet on the document level. TextRank weights the graph according the PageRank scheme, which is different from our method (we used the normalized Laplacian as commonly used in graph-based machine learning [50]). The PageRank scheme will be further exploited in future work. As it is infeasible to directly compare our method with TextRank, in the subsequent experiments we compare our method to more relevant and state-of-the-art personalized cross-language search methods.

4. Latent semantic-based adaptation

In this section, we introduce how we obtain the scores from non-personalized rankers and personalized rankers. As previously mentioned, in our method, we can simultaneously process multiple adaptation scores without the complex parameter tuning procedures. We present the calculation of $S_{term}(q, d)$, $S_{latent}(q, d)$ and $S_{user}(q, d)$.

Instead of using a lexical matching method and a method exploring external semantics, we adopt latent-semantic based adaptation. We extend previous work by using the LDA model [49]. Given a set of documents, we are trying to measure 1) the lexical distance between the initial query and a document (i.e.$S_{term}(q, d)$); 2) the semantic distance between the initial query and a document (i.e. $S_{latent}(q, d)$); and 3) the semantic distance between the user-enhanced query and a document (i.e. $S_{user}(q, d)$).

4.1. Non-personalized adaptation

We calculate the lexical distance from a non-commutative measure of the difference between two probability distributions. The distance used in this approach is the Kullback-Leibler (KL) divergence [23]. In terms of text sequences (either queries or documents), the probability distribution can be regarded as a probabilistic language model M_d or M_q from each document d or each query q. In other words, it assumes that there is an underlying language model which "generates" a term (sequence) [29]. The unigram language model is utilized here.

There are several ways to estimate the probabilities. Let $g(t \in d)$ denote the number of times the term t occurs in a document d (the same idea can be used on a query). The Maximum-likelihood estimation (MLE) of t with respect to d is defined as:

$$MLE_d t \overset{\text{def}}{=} \frac{g(t \in d)}{\sum_{t'} g(t' \in d)}$$

Previous work in language-model-based information retrieval [45] advocates the use of a Dirichlet-smoothed estimation:

$$DIR_d t \overset{\text{def}}{=} \frac{g(t \in d) + \mu \cdot MLE_t D}{\sum_{t'} g(t' \in d) + \mu}$$

where a smoothing parameter μ controls the degree of reliance on relative frequencies in the document corpus rather than on the counts in d. The $S_{term}(q, d)$ that we choose to use computes the KL divergence between the $MLE_q t$ and a modified version of $DIR_d t$ as in [45].

With respect to the semantic distance, we estimate the probability of a document d generating t, using a mixture model LDA. It uses a convex combination of a set of component distributions to model observations. In this model, a term t is generated from a convex combination of some hidden topics z:

$$LDA_d(t) = \sum_{z=1}^{k} p(t|z)p(z|d) \tag{9}$$

where each mixture model $p(t|z)$ is a multinomial distribution over terms that correspond to one of the latent topics z. Then the distance between a query and a document based on this model can be obtained. The semantic-based method we use here adopts the KL divergence between the query terms and document terms to compute a $S_{latent}(q, d)$ score:

$$S_{latent}(q, d) = -D(MLE_q(\cdot)||LDA_d(\cdot)) \tag{10}$$

This method also has the property of length-normalization to ameliorate long document bias problems.

Table 2
Co-occurrence based query enhancement algorithm.

```
Input: The initial query $q$,
  The user model.
Output: The enhanced query $q'$.
Let $S$ be the set of keywords in the user model that could potentially be added to an input query $q$.
for each term $t$ of $q$ do
    $S \leftarrow S \cup Top(t)$ where $Top(t)$ contains the top terms with the closest relationship to $t$ (obtained from co-occurrence statistics)
for each term $t_j$ in $S$ do
    $COOCCUR(t_j) = \prod_{t_i \in q} (0.01 + COS(t_i,\ t_j))$ //0.01 is to avoid zero value without adding bias
Select top $\gamma$ terms of $S$ with highest $COOCCUR(t_j)$ scores with $q$ to form $q'$.
```

4.2. Personalized adaptation

To produce personalized adaptation scores, we compute the KL divergence between documents and user enhanced query. The user enhanced query is computed by using the user model to obtain $S_{user}(q, d)$. We describe the procedure for acquiring the user enhanced query, and the user model generation here.

A user model is learned from the user's historical usage information. Here a user is assumed to perform daily searches in one language, and occasionally s/he wants to search for information in different languages. So the user model stores terms which represent the user's interests in the main language in which s/he performs daily searches. The whole model is monolingual and there is no human interaction in target language. 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. Note here the query and the documents are in the same language. Then the documents are processed to extract the terms that are most representative of them. To define the representative terms, frequency-based methods can be applied. The extracted terms along with the query terms are subsequently assigned weights accordingly.

As the simplest possible measures, term frequency (*tf*) and inverse document frequency (*idf*) have the advantage of being very fast to compute. Previous experiments with small datasets have shown them to yield very good results [21]. We use the following equation to obtain the weight:

$$TFIDF_d t = \frac{g(t \in d)}{max_{t'} g(t' \in d)} \cdot log \frac{|D|}{df_t} \tag{11}$$

where $|D|$ is the total number of documents and df_t denotes document frequency of the term t in the corpus.

We also consider another user model generation technique by using the *BM25* scheme, which is defined as:

$$BM25_d t = \sum_t weight_t \frac{(k_1+1)g(t \in d)}{K + g(t \in d)} \cdot \frac{(k_3+1)g(t \in d)}{k_3 + g(t \in d)} \tag{12}$$

where $weight_t$ is the Robertson/Sparck Jones weight of t and is defined as:

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

K is defined as:

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

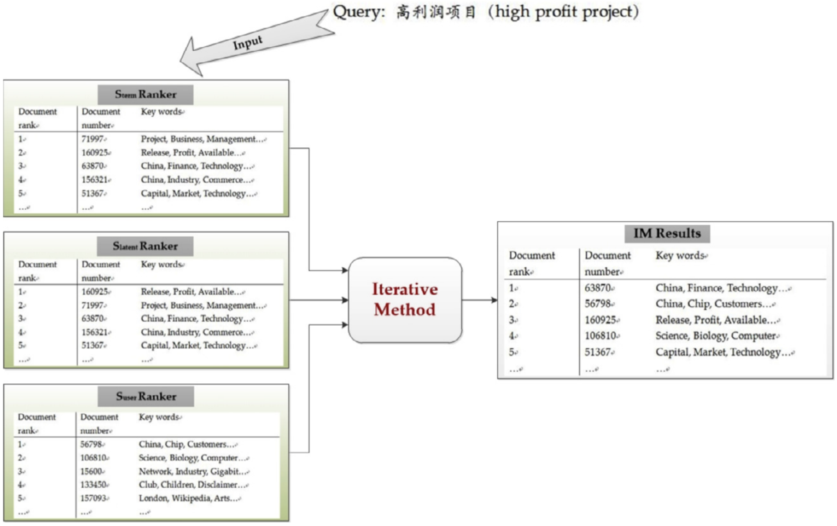
k_1, b, and k_3 are parameters (set to 1.2, 0.75, and 7 respectively), $|d|$ represents document length and $avg|d|$ stands for average document length.

We now present the procedure to obtain the user-enhanced query from the initial query. These techniques add terms coming from the user model to the original query. The technique exploits co-occurrence of query terms with the profile keywords. 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 enhancement algorithm is presented in Table 2. It is worth noting that the unclicked documents might have a negative effect on term re-weighting, an investigation of this remains as future work.

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

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

The reason we use the document frequency in the calculation is that we are focused on the collocation within a document, so the individual frequency of one term is that term's document frequency. Previous researches show the effect of this approach, for example [20].



S_{term} Ranker

Document rank	Document number	Key words
1	71997	Project, Business, Management…
2	160925	Release, Profit, Available…
3	63870	China, Finance, Technology…
4	156321	China, Industry, Commerce…
5	51367	Capital, Market, Technology…
…	…	…

S_{latent} Ranker

Document rank	Document number	Key words
1	160925	Release, Profit, Available…
2	71997	Project, Business, Management…
3	63870	China, Finance, Technology…
4	156321	China, Industry, Commerce…
5	51367	Capital, Market, Technology…
…	…	…

S_{user} Ranker

Document rank	Document number	Key words
1	56798	China, Chip, Customers…
2	106810	Science, Biology, Computer…
3	15600	Network, Industry, Gigabit…
4	133450	Club, Children, Disclaimer…
5	157093	London, Wikipedia, Arts…
…	…	…

IM Results

Document rank	Document number	Key words
1	63870	China, Finance, Technology…
2	56798	China, Chip, Customers…
3	160925	Release, Profit, Available…
4	106810	Science, Biology, Computer
5	51367	Capital, Market, Technology…
…	…	…

Fig. 1. An illustrative example.

We then adopt the KL divergence between the enhanced query terms and document terms to compute a $S_{user}(q, d)$ score, where q' represents the user's enhanced query:

$$S_{user}(q,d) = -D(MLE_{q'}(\cdot)||LDA_d(\cdot)) \tag{14}$$

4.3. An illustrative example

To illustrate the advantage of our proposed iterative method, we give an example in Fig. 1 from the log-based test collection, using different methods. As we can see, for the query "high profit project" (note the query is in Chinese, we provide a term-by-term translation here), the S_{term} ranker (correspondent to lexical-based retrieval model) can only rank retrieval results according to lexical matching (Note only representative terms are shown for each document, and only the first 5 results are listed). The S_{latent} ranker (correspondent to semantic-based retrieval like the one described in Section 4.1) only changes the order of the first two results to reflect the semantic matching. The S_{user} ranker (correspondent to personalized retrieval like we described in Section 4.2) performs differently, by adding several results that do not appear in the list of the other two base rankers, but which are quite relevant to the user. However, it also introduces some noise, by retrieving documents that are more relevant to the user model, but not to the issued query. By fusing the three sets of results together, our iterative method works quite well, by considering both the issued query and the user's interests. By checking the actual web resources the particular user clicked, we found that indeed s/he was looking for high profit projects like chips and biology that could be invested in China.

5. Experiments and results

In the following section we describe experiments which have been designed to evaluate the proposed method. We start the section by discussing the experimental settings, and then we present and analyze the results. This evaluation focuses on the following thematically related questions:

i. *Is the proposed iterative refinement method an improvement over classical non-personalized and linear-combination-based personalized techniques for personalized results adaptation in the context of cross-language search?*
ii. *What is the validity of results obtained on the semi-automatically created test collection, created by our proposed evaluation methodology?*
iii. *Can a simple user model provide effective results in personalized cross-language search?*

5.1. Experimental settings

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 cross-

[2] http://www.clef-initiative.eu/.
[3] http://research.nii.ac.jp/ntcir/index-en.html.

language search. 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 emails and desktop documents, are not easy to acquire. Information stored by a search engine provider is normally unavailable to researchers outside the organization.

Motivated by Vicente-López *et al.* [38], we created a test collection to semi-automatically evaluate the proposed approaches. In Vicente-López et al.'s evaluation framework, there are four main components specified. Document collections are 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 cross-language search, as cross-language relevancy is more difficult to obtain. Hence, we develop a cross-language evaluation based on the above procedure.

A Wikipedia database consisting of documents in Chinese and English was used to construct the test collection. Only those articles that are connected via cross-language links between the two Wikipedia databases were selected. A snapshot was obtained on the 14/08/2014, which contained an aligned collection of 158,037 articles in the two languages. Some of the articles are written independently and by different authors, and some of the articles are direct translations of each other. It is worth noting that we discard non-text features like the structure of Wikipedia articles, language use and connectivity etc. to avoid any possible bias caused by Wikipedia. We only consider plain text. We also note that Wikipedia articles are different from Web documents, however, we only use this collection to illustrate the effectiveness of our method in comparison to existing methods, in the absence of real log data. This has been confirmed as a valid approach in the literature, such as in Vicente-López et al. [38] and Azzopardi et al.'s [2] work. We show later in the paper that when compared to results obtained using a log-based collection, we found no significant differences.

We assume that a user performs daily searches in their native language (referred to as the source language, in our case, Chinese), and occasionally s/he wants to search in another language (target language, i.e. English). Hence cross-language search is performed by firstly translating Chinese queries into English queries using a translation mechanism, then retrieving English documents from the test collection. In order to simulate users, Wikipedia articles are first clustered into several categories to represent user interests. This is done in the source language (Chinese), which is assumed to be user's daily search language. The Chinese collection is first grouped into 1362 clusters. Each of the clusters could be used to generate a user profile. This is possible because, in theory, clusters represent different areas of interests. 75% of documents inside each cluster are chosen to build user profiles, while the remaining 25% are left for testing. Note, the number of clusters is not fixed to a value as when clustering the category of the document is also considered.

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

The next question is to generate queries and relevance judgments. This is done by simulating *cross-language known-item search* [2]. It provides precise semantics and removes the burden to build queries and relevance judgments. The search process assumes that only one document is relevant for a specific query. Here that task is then reduced to find the correct Wikipedia article in the target language with a query provided in the source language. In order to produce more accurate and realistic relevance judgments, we include a human in the loop. 40 undergraduate and postgraduate students manually checked the results retrieved by the training queries. They were instructed to judge each cross-language item as relevant or not by assuming his chosen user model (i.e. the cluster). They were assigned similar numbers of clusters to judge. If s/he feels that most of the documents are not relevant, then that cluster will be marked. Each cluster has been reviewed by at least three subjects, if two of them mark the cluster, then it will be discarded. This process filtered out 340 clusters and left 1022 clusters for the final evaluation.

The queries are also automatically generated according to Azzopardi et al.'s work [2]. 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 procedure described in Table 3 is used to generate a much shorter query analog to a real web query. As we mainly use it to generate queries, according to Azzopardi et al.'s work [2], as δ tends to zero, the user's recollection of the original document improves. If $\delta = 1$, then the user knows the document exists but they have no idea as to which terms appear in the document (and randomly selects query terms). $P(t_i|d^C)$ is calculated as:

$$P\left(t_i|d^C\right) = \frac{g\left(t_i \in d^C\right) \cdot log\frac{|D|}{df_{t_i}}}{\sum_{t_j}\left(g\left(t_j \in d^C\right) \cdot log\frac{|D|}{df_{t_j}}\right)} \tag{15}$$

In order to verify the validity of the semi-automatically generated corpus, we used the second test collection which is built upon a commercial search log released for the NTCIR-9 Intent task. This corpus contains 135.4 million Web pages crawled from 5.3 million Chinese Web sites during 2008. The total uncompressed storage size of this collection is approximately 5 TB. The collection contains text only. Both the search log and the texts are fully available from the task organizer

Table 3
Procedure for generating queries.

```
Input: The bilingual Wikipedia document pair ($d^C$, $d^E$).
Output: The generated query $q^C$.
1: Pick the Wikipedia document $d^E$and its aligned document $d^C$in another language
2: Initialize an empty query $q^C$
3: Choose query length $L$ with the Poisson distribution $poi(L)$. The mean is set to 3, which is closest to the average
   length of a real web query according to the statistics of the log data
4: for each term $t_i$ in $d^C$ do
5: $Score(t_i) = (1-\delta)\cdot P(t_i|d^C) + \delta\cdot P(t_i|Collection^C)$
6: Rank all terms from the document $d^C$ based on the scores computed at step 5
7: Select top $L$ terms with highest scores to form $q^C$
```

or directly from the Sogou Company.[4] The search log contains all of the click-through data collected by a Chinese search engine over the web pages during June 2008. In this log, each click event is represented by a single line of tab-separated data containing the following fields: the ID of the user performing the search; the user's query; the ranking of the clicked URL; the ordering of the user click, and the URL that was clicked. Note, in this experiment, we assume that a click event on a specific URL indicates that the clicked Web page is relevant to the submitted query, i.e. we are discounting accidental/erroneous click events. This will obviously introduce some noise to the relevance judgment process as in all log-based user studies [17]. In future, we will investigate this issue further by recruiting and monitoring real users' behavior.

We chose users from the log who had submitted both Chinese and English queries. In particular, they will need to have English results judged for the corresponding Chinese queries. This process is to find out the Chinese-English cross-language search events inside the log. In total we found 212 users within the chosen date range. The Chinese queries with English results clicked are chosen as the test data and the remaining queries and results are left as training data to build user models.

The English terms were processed in the usual way, i.e. down-casing the alphabetic characters, removing the stop words and stemming words using the Porter stemmer. Chinese documents were segmented using a freely available analyzer.[5] No other filtering is conducted. We follow Cao et al.'s approach [11] to translate the original and enhanced queries. The translation probabilities for a query term are obtained with the GIZA++ toolkit, which extracts a statistical translation model from the bilingual dictionary, considered as a parallel corpus. As in their paper, this model can produce better cross-language search retrieval performance than many state-of-the-art dictionary-based cross-language search approaches. The bilingual dictionary is combined from the LDC Chinese-English dictionary[6] and a bilingual lexicon generated from a parallel corpus.[7] OOV terms are mined from the Web following the approach in [48]. As the primary focus of the current paper is the results adaptation rather than query adaptation, we simply adopt one representation method for query translation. We understand that query translation is particularly important for cross-language search. We plan to fully exploit the use of modern statistical machine translation systems for query translation in cross-language web search such as multiple representations [36] and compare the differences in terms of evaluation effectiveness when combining with our results adaptation method. All the information retrieval experiments were performed using the Terrier[8] platform.

5.2. Evaluation metrics

The following evaluation metrics were chosen to measure the effectiveness of the various approaches: mean reciprocal rank (MRR), normalized discounted cumulative gain (NDCG), the precision of the top 1 documents (P@1) and the precision of the top 5 documents (P@5). The first three measurements are commonly used to evaluate personalized search algorithms [12,17] while the last one is useful for evaluating known-item search [2]. The four metrics were calculated for each user and the mean of all the values was calculated, so that the average performance over test users could be computed. Statistically-significant differences in performance were determined using a paired t-test at a confidence level of 95%. We retrieve 1000 documents in each retrieval.

5.3. Experimental runs

The proposed approach is applied to personalized cross-language results adaptation. We evaluate our proposed models and compare with several state-of-the-art non-personalized and personalized methods as follows.

[4] The website for the task is at http://www.thuir.cn/intent/ntcir9, and the dataset is also available at http://www.sogou.com/labs/dl/t-e.html.

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

[6] Available from htttps://catalog.ldc.upenn.edu/LDC2002L27.

[7] Available from htttps://catalog.ldc.upenn.edu/LDC2014T04.

[8] http://www.terrier.org.

Table 4
Experimental results for the Wikipedia-based collection, statistically significant differences between the method and LM, LMRM, QE, LEXICAL, ODP and LDA are indicated by *l, r, e, x, o, a* respectively.

	TFIDF				BM25			
	P1	P5	NDCG	MRR	P1	P5	NDCG	MRR
LM	0.1667	0.0583	0.2750	0.2275	0.1667	0.0583	0.2750	0.2275
LMRM	0.1875^{l}	0.0583	0.2912^{l}	0.2443	0.1875^{l}	0.0625	0.2917^{l}	0.2443
QE	0.1858^{l}	0.0595^{l}	0.2816^{l}	0.2337^{l}	0.1958^{l}	0.0595^{l}	0.2838^{l}	0.2412^{l}
LEXICAL	$0.2083^{l,r,e}$	$0.0625^{l,r,e}$	$0.2917^{l,r,e}$	$0.2518^{l,r,e}$	$0.2083^{l,r,e}$	$0.0625^{l,r,e}$	$0.3024^{l,r,e}$	$0.2664^{l,r,e}$
ODP	$0.2083^{l,r,e}$	$0.0625^{l,r,e}$	$0.3049^{l,r}_{e,x}$	$0.2658^{l,r}_{e,x}$	$0.2083^{l,r,e}$	$0.0667^{l,r,e}$	$0.3077^{l,r}_{e,x}$	$0.2687^{l,r}_{e,x}$
LDA	$0.2083^{l,r,e}$	$0.0625^{l,r,e}$	$0.3077^{l,r}_{e,x,o}$	$0.2769^{l,r}_{e,x,o}$	$0.2292^{l,r}_{e,x,o}$	$0.0625^{l,r,e}$	$0.3123^{l,r}_{e,x,o}$	$0.2687^{l,r}_{e,x}$
IM	$0.2292^{l,r,e}_{x,o,a}$	$0.0667^{l,r,e}_{x,o,a}$	$0.3135^{l,r,e}_{x,o,a}$	$0.2781^{l,r,e}_{x,o,a}$	$0.2500^{l,r,e}_{x,o,a}$	$0.0708^{l,r,e}_{x,o,a}$	$0.3194^{l,r,e}_{x,o,a}$	$0.2861^{l,r,e}_{x,o,a}$

LM A popular and quite robust language model retrieval method which has previously shown good results, as in [45]. It computes the KL divergence between the $MLE_q w$ (Eq. (7)) and a modified version of $DIR_d w$ (Eq. (8)). It can be viewed as a lexical-based matching method.

LMRM A non-personalized relevance model involves pseudo-relevance feedback in the language model as in [22]. Readers can refer to their paper for details. This is another lexical-based method.

QE This method uses the user model generation techniques and the query expansion technique introduced in Section 4.2 for personalization (i.e. use Eq. (14) to produce results). It can be viewed as a baseline to compare query adaptation and results adaptation. It is also used here to compare to our iterative method.

LEXICAL This is the only method in the literature dealing with personalized multilingual results adaptation [16]. In order to do a fair comparison, we also implemented multiple vectors in each set representing multiple clusters of interests associated with the language. In cases where only single language results are returned, we use the Bing translator to translate all results.

ODP This personalized method maps the documents and user models onto the ODP categories as in [13]. They used a distance function to calculate the distances between user models and ODP categories as well as documents retrieved by a user query and the ODP categories. Then the similarities between documents and user models can be re-calculated so as to re-rank the initial results. We used 1171 second-level categories in our implementation of the method.

LDA In this personalized method, we linearly fuse the three lists of scores $S_{term}(q, d)$, $S_{latent}(q, d)$ and $S_{user}(q, d)$ as in many methods in literature: $(1 - \lambda_1 - \lambda_2)S_{term}(q, d) + \lambda_1 S_{latent}(q, d) + \lambda_2 S_{user}(q, d)$. λ_1, λ_2 are parameters that can be tuned with $\lambda_1 + \lambda_2 = 1$ meaning no results adaptation is performed. We tune the parameter to maximize the MRR in one collection and use it in all the test collections. $S_{latent}(q, d)$ and $S_{user}(q, d)$ are calculated according to the procedures described in Section 4.1 and 4.2, respectively. In previous work [49], we have proven that $S_{latent}(q, d)$ is superior than using the LDA model described in [41].

IM Our proposed iterative method, which assumes that results retrieved by non-personalized rankers and personalized rankers mutually reinforce each other, rather than being linearly combined

With regard to the parameters, α controls the fraction of score information that a document receives from its neighbors. If we want a document to retain more score information from its initial rankers, α will be set to a relatively small number, otherwise it will be set to a larger number. In our experiments, as in other label propagation methods [46,50], we set α to 0.99, which means we want a document to absorb more information from its neighborhood. As described in previous work [8,25], the retrieval weights are affected by the relations between the text and its neighboring, which may not be trivial. We will show later on this parameter has a relatively small effect on the final results, only setting it too small will affect the performance (see section below). γ is empirically set to 50 for query enhancement. According to Azzopardi et al. [2] and through our own experimentation, we set δ to 0.2.

5.4. Experimental results

The experimental results for the results adaptation in the Wikipedia-based collection are presented in Table 4. **TFIDF** and **BM25** represent user model generation techniques. As illustrated by the results, the **LM** model performed consistently poorly for all evaluation metrics. **LMRM** provides an improvement over the simple query term matching, this result is consistent with the previous study and confirms that the relevance model has certain benefits for non-personalized search. However, all personalized runs outperform these two non-personalized baselines except the **QE** method. The performance of the query expansion-based method is even lower than the **LMRM** method in most of evaluation metrics. This demonstrates that simple query expansion methods will not produce good personalization results, more sophisticated methods are be needed. Our proposed **IM** method performs especially well. It achieved statistically significant improvements over the top performing baselines, including the external semantic (marked as *o*) and latent semantic (marked as *a*) based models. Notably, it scored a 10.03% improvement over the **LEXICAL, ODP** and **LDA** models and an 18.93% improvement over the **QE** model measured by **P1** and using the **TFIDF** user model generation technique. This improvement is significant for the Wikipedia known-item search. The results adaptation models **LEXICAL, ODP** and **LDA** have also achieved good improvements

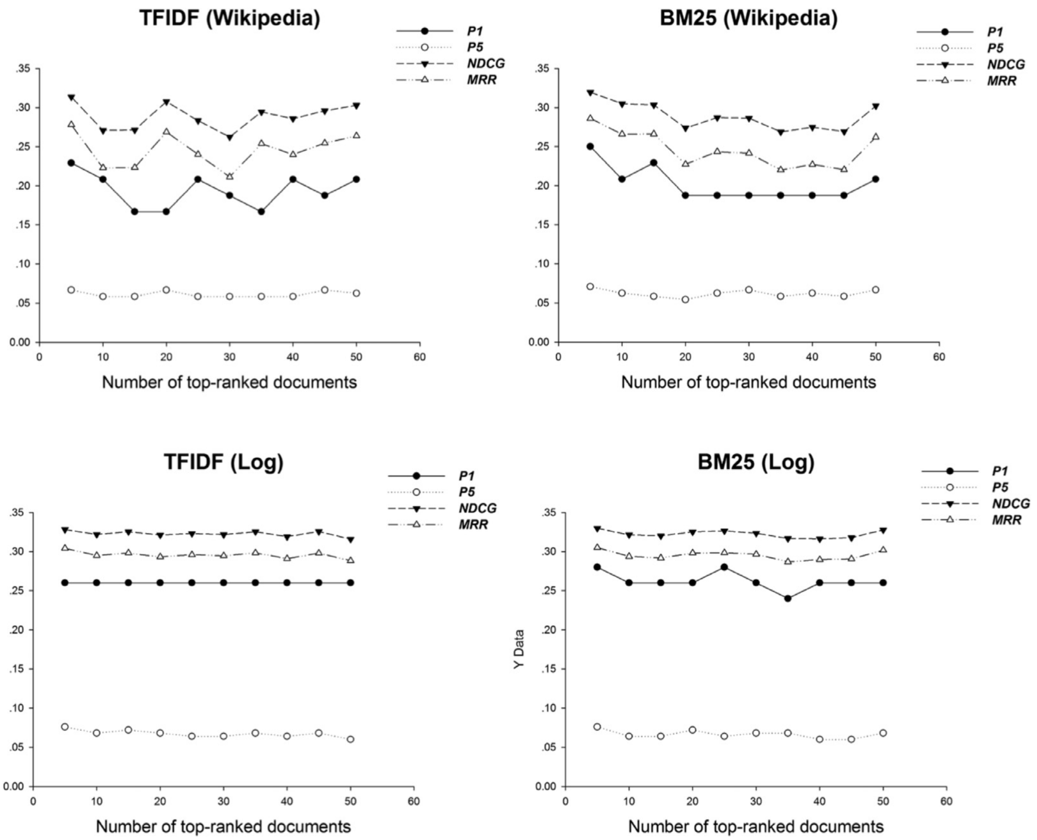


Fig. 2. Comparison of personalized query expansion techniques.

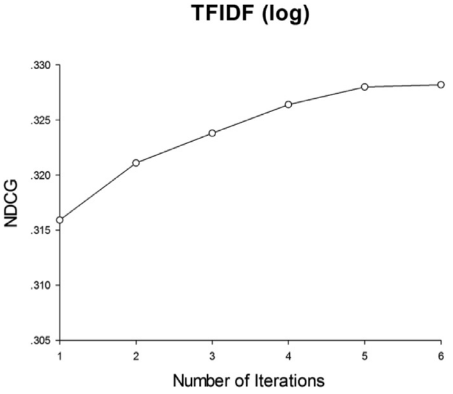


Fig. 3. Performance comparison with different number of iterations.

over the non-personalized models. **ODP** and **LDA** both perform better than the **LEXICAL** model. This confirms that simple lexical match techniques may miss semantic information when ranking. The external and latent semantic-based methods show similar performance, while the latent semantic-based models are the most straightforward to implement. Even with the low performance of the query expansion-based method, including it into our iterative method will bring significant improvements. This result also shows the power of our proposed method.

Another observation from the results is that using monolingual user models can enhance personalized cross-language search. We notice that personalized results adaptation techniques produce very good results using the two types of user models, which are always better than the corresponding results of all the non-personalized strategies with statistically significant results. The performance improvements are quite stable for all evaluated methods.

To confirm the validity of the results obtained by using our semi-automatic methodology, we repeat the experiments using the log-based test collection. Similar results are obtained, as shown in Table 5. The **IM** method is still the best strategy

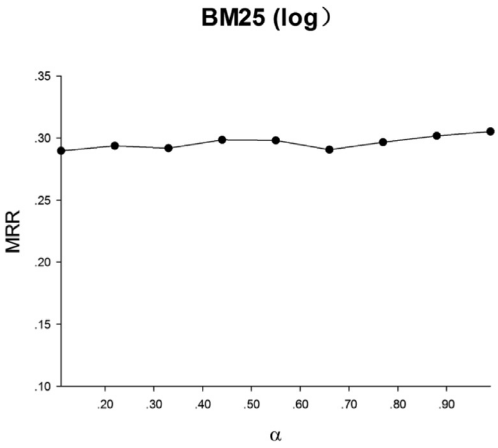


Fig. 4. Performance of different values of parameter α.

Table 5
Experimental results for the log-based collection, statistically significant differences between the method and LM, LMRM, QE, LEXICAL, ODP and LDA are indicated by *l, r, e, x, o, a* respectively.

	TFIDF				BM25			
	P1	P5	NDCG	MRR	P1	P5	NDCG	MRR
LM	0.2000	0.0600	0.2844	0.2448	0.2000	0.0600	0.2844	0.2448
LMRM	0.2200	0.0680^{l}	0.2971^{l}	0.2713^{l}	0.2200	0.0680^{l}	0.3042^{l}	0.2631^{l}
QE	0.2100	0.0640^{l}	0.2901^{l}	0.2573^{l}	0.2100	0.0640^{l}	0.2916^{l}	0.2504^{l}
LEXICAL	0.2200	0.0680^{l}	$0.3042^{l,r,e}$	$0.2820^{l,r,e}$	0.2200	0.0680^{l}	$0.3130^{l,r,e}$	$0.2713^{l,r,e}$
ODP	$0.2400^{l,r}_{e,x}$	0.0680^{l}	$0.3159^{l,r}_{e,x}$	$0.2883^{l,r}_{e,x}$	$0.2600^{l,r}_{e,x}$	$0.0720^{l,r}_{e,x}$	$0.3169^{l,r}_{e,x}$	$0.2888^{l,r}_{e,x}$
LDA	$0.2400^{l,r}_{e,x}$	0.0680^{l}	$0.3202^{l,r}_{e,x,o}$	$0.2859^{l,r}_{e,x}$	$0.2600^{l,r}_{e,x}$	0.0680^{l}	$0.3203^{l,r}_{e,x,o}$	$0.2921^{l,r}_{e,x,o}$
IM	$0.2600^{l,r,e}_{x,o,a}$	$0.0760^{l,r,e}_{x,o,a}$	$0.3282^{l,r,e}_{x,o,a}$	$0.3040^{l,r,e}_{x,o,a}$	$0.2800^{l,r,e}_{x,o,a}$	$0.0760^{l,r,e}_{x,o,a}$	$0.3298^{l,r,e}_{x,o,a}$	$0.3052^{l,r,e}_{x,o,a}$

for personalized results adaptation, better than the **QE, LEXICAL, ODP** and **LDA** models. We also notice that the performance improvements in the log-based test collection are also quite stable with the highest improvement reaching 11.76% in terms of the **IM** method compared to the **ODP** and **LDA** baselines using the **P5** metric and the **TFIDF** user model generation technique. The improvements using the NDCG metric are not as obvious as those using the precision-based metrics. This is perhaps caused by the immediate non-relevant documents after the first few relevant documents. More experiments could be performed in the future to further analyze these results. All the results adaptation oriented personalized approaches work better than the non-personalized approaches by a large margin with statistically significant results. The similar performance observed in both test collections demonstrates the success of the semi-automatically constructed test collection. It provides an interesting alternative when real user data is not available.

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

Next we examine the effect on the performance when we vary the number of top-ranked documents that we choose to give scores for each ranker in our method (i.e. parameter β). We vary the number of topics in both methods from 5 to 50, the results are shown in Fig. 2. As can be seen from the figure, the highest performance is reached when the number of top-ranked documents is 5 in both collections. When the number varies, the performance also changes quickly. This means that a small number of topics can capture the semantics exhibited in the user model. This also shows the importance of the parameter β on the results. It performs more stably in the log-based test collection than in the Wikipedia-based test collection, this may be caused by the greater topic coverage in the Wikipedia articles.

In Fig. 3 we present the NDCG results with different numbers of iterations in the iterative process of our proposed method **IM**. We only provide results in the log-based test collection by using the TFIDF user model generation technique as it demonstrates similar behavior in other situations. From the figure we can see that the performance consistently improves when the number of iterations increases. However, the curve converges very quickly. It can be seen that the improvement becomes very limited after 5 iterations. So that in our experiments we set the iteration number to 6 in both test collections.

In Fig. 4 we present the MRR results with different values of α (from 0.11 to 0.99) in our proposed method **IM** as this parameter may have non-trivial effects on the results. We only provide results in the log-based test collection by using the

BM25 user model generation technique as it demonstrates similar behavior in other situations. From the figure we can see that the highest performance is obtained when the value reaches 0.99. In fact, we found that this parameter has a relatively small effect on the final results, only setting it to be too small will significantly affect the performance.

6. Conclusions

In this paper, we propose a novel iterative method based on document associations obtained from an initial ranker. The method assumes that results retrieved by non-personalized rankers and personalized rankers mutually reinforce each other rather than being linearly combined. This method, combined with latent-semantic techniques, produces very good results for personalized results adaptation when applied in the context of cross-language search. In addition to the evaluation of the personalized result adaptation technique, we also study the effects of user model generation. The results show that a simple vector-space method is sufficient for building user models in this context. We also confirmed that user models generated from historical usage information in one language can enhance search in another language. Furthermore, the experimental results also prove that the semi-automatically constructed test collection proposed in this paper can be used as an alternative approach to evaluation in the absence of real world datasets.

We can now comfortably answer the three research questions we proposed at the beginning of Section 5:

i. *Is the proposed iterative refinement method an improvement over classical non-personalized and linear-combination-based personalized techniques for personalized results adaptation in the context of cross-language search?*
The answer to this question is clearly, yes. As we can see from the experimental results, our proposed method outperforms classical non-personalized search approaches, including language models and relevance models. It can also produce better results than several state-of-art personalized methods for cross-language web search using different user model representations. These methods include those exploiting lexical matching, and external and latent semantics to rerank search results. Moreover, we prove that our iterative refinement method, which assumes that results retrieved by non-personalized rankers and personalized rankers mutually reinforce each other, works significantly better than linear-combination-based personalized approaches. This finding has two implications: 1) it provides an opportunity to add effective personalization functions to current search systems; 2) it is possible to go beyond linear-combination-based approaches for personalized results adaptation, through the cross-language search scenario;

ii. *What is the validity of results obtained on the semi-automatically created test collection created by our proposed evaluation methodology?*
This validity is verified through our experiments. We notice that the performance improvements achieved on the log-based test collection are very close to those achieved on the semi-automatically created collection. All personalized approaches work better than the non-personalized approaches by a large margin with statistically significant results. Similar performance is observed in both test collections, which demonstrates that the semi-automatically constructed test collection can be used as an alternative dataset for evaluation in the absence of available real-world datasets. This finding provides an evaluation methodology to help lower the common high barrier to the evaluation of cross-language personalized systems.

iii. *Can a simple user model provide effective results in personalized cross-language search?*
In the experiments detailed in this paper, both user model representation techniques demonstrated slightly different performance in generating user models for use in personalized search. The results show that in terms of user model generation, simple user model generation techniques work well. This finding provides a simple yet effective monolingual-based user model representation technique to be used in personalized cross-language search.

In this paper, personalization strategies are investigated independently. In future work they could be integrated in a more unified way to enhance personalized cross-language search. We also plan to investigate more user model representation techniques for personalized search, especially those exploiting multilingual facilities. The semi-automatic test collection generation process can be further improved by exploiting more external corpora. In the current paper we only test the English-Chinese language pair for our method, this is for the following two reasons: first, the search log data we use mainly contains users from China, they can read English but have difficulty in formulating English queries, therefore we only test the English-Chinese direction. Second, we want the experiments to remain consistent across both test collections, hence we keep the same language pairs. We are prepared to test more language pairs and directions in the very near future. Last but not least, our method is designed for personalized search and can be applied in many contexts. As such, we are also prepared to test this approach in application domains other than the cross-language search.

Acknowledgments

The work described in this paper is supported by Hunan Provincial Natural Science Foundation of China under Grant No. 2017JJ2101, Scientific Research Fund of Hunan Provincial Education Department of China under Grant No. 16K030, National Natural Science Foundation of China under Project No. 61300129, No. 61572187 and No. 61272063. 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 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

[1] J.A. Aslam, M. Montague, Models for metasearch, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, Louisiana, USA, ACM, 2001, pp. 276–284. http://dx.doi.org/10.1145/383952.384007.

[2] L. Azzopardi, M.D. Rijke, K. Balog, Building simulated queries for known-item topics: an analysis using six European languages, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, ACM, 2007, pp. 455–462. http://dx.doi.org/10.1145/1277741.1277820.

[3] M. Bayomi, K. Levacher, M.R. Ghorab, P. Lavin, A. O'connor, S. Lawless, Towards evaluating the impact of anaphora resolution on text summarisation from a human perspective, in: E. Métais, F. Meziane, M. Saraee, V. Sugumaran, S. Vadera (Eds.), Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings, Springer International Publishing, Cham, 2016, pp. 187–199. http://dx.doi.org/10.1007/978-3-319-41754-7_16.

[4] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[5] M.R. Bouadjenek, H. Hacid, M. Bouzeghoub, Sopra: a new social personalized ranking function for improving web search, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2013, pp. 861–864.

[6] M.R. Bouadjenek, H. Hacid, M. Bouzeghoub, Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms, Inf. Syst. 56 (2016) 1–18.

[7] M.R. Bouadjenek, H. Hacid, M. Bouzeghoub, A. Vakali, Using social annotations to enhance document representation for personalized search, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2013, pp. 1049–1052.

[8] D. Cai, Q. Mei, J. Han, C. Zhai, Modeling hidden topics on document manifold, in: Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, California, USA, ACM, 2008, pp. 911–920. http://dx.doi.org/10.1145/1458082.1458202.

[9] F. Cai, S. Liang, M.D. Rijke, Personalized document re-ranking based on Bayesian probabilistic matrix factorization, in: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast, Queensland, Australia, ACM, 2014, pp. 835–838. http://dx.doi.org/10.1145/2600428.2609453.

[10] Y. Cai, Q. Li, Personalized search by tag-based user profile and resource profile in collaborative tagging systems, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, ON, Canada, ACM, 2010, pp. 969–978. http://dx.doi.org/10.1145/1871437.1871561.

[11] G. Cao, J. Gao, J.-Y. Nie, J. Bai, Extending query translation to cross-language query expansion with Markov chain models, in: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, Lisbon, Portugal, ACM, 2007, pp. 351–360. http://dx.doi.org/10.1145/1321440.1321491.

[12] P.-A. Chirita, C.S. Firan, W. Nejdl, Personalized query expansion for the web, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, ACM, 2007, pp. 7–14. http://dx.doi.org/10.1145/1277741.1277746.

[13] P.A. Chirita, W. Nejdl, R. Paiu, C. Kohlschutter, Using ODP metadata to personalize search, in: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, ACM, 2005, pp. 178–185. http://dx.doi.org/10.1145/1076034.1076067.

[14] M. Claypool, D. Brown, P. Le, M. Waseda, Inferring user interest, IEEE Internet Comput. 5 (6) (2001) 32–39. http://dx.doi.org/10.1109/4236.968829.

[15] J.M. Conroy, S.T. Davis, J. Kubina, Y.-K. Liu, D.P. O'leary, J.D. Schlesinger, Multilingual summarization: dimensionality reduction and a step towards optimal term coverage, in: MultiLing Workshop, 2013, pp. 55–63.

[16] M.R. Ghorab, D. Zhou, S. Lawless, V. Wade, Multilingual user modeling for personalized *re*-ranking of multilingual web search results, in: CEUR Workshop Proceeding of UMAP 2012, Montreal, Canada, 2012, pp. 1–4.

[17] M.R. Ghorab, D. Zhou, A. O'connor, V. Wade, Personalised information retrieval: survey and classification, User Model. User-Adapted Interact. 23 (4) (2013) 381–443. http://dx.doi.org/10.1007/s11257-012-9124-1.

[18] M.R. Ghorab, D. Zhou, B. Steichen, V. Wade, Towards multilingual user models for Personalized Multilingual Information Retrieval, in: Proceedings of the First Workshop on Personalised Multilingual Hypertext Retrieval, Eindhoven, Netherlands, ACM, 2011, pp. 42–49. http://dx.doi.org/10.1145/2047403.2047411.

[19] H.B. Hashemi, A. Shakery, Mining a Persian-English comparable corpus for cross-language information retrieval, Inf. Process. Manage. 50 (2) (2014) 384–398. (/) http://dx.doi.org/10.1016/j.ipm.2013.10.002 .

[20] M.-C. Kim, K.-S. Choi, A comparison of collocation-based similarity measures in query expansion, Inf. Process. Manage. 35 (1) (1999) 19–30. (1999/01/01/) http://dx.doi.org/http://dx.doi.org/10.1016/S0306-4573(98)00040-5 .

[21] R.R. Larson, Introduction to information retrieval, J. Am. Soc. Inf. Sci. Technol. 61 (4) (2010) 852–853. http://dx.doi.org/10.1002/asi.21234.

[22] V. Lavrenko, W.B. Croft, Relevance based language models, in: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2001, pp. 120–127.

[23] J. Lin, Divergence measures based on the Shannon entropy, Inf. Theory, IEEE Trans. 37 (1) (1991) 145–151.

[24] Y.-T. Liu, T.-Y. Liu, T. Qin, Z.-M. Ma, H. Li, Supervised rank aggregation, in: Proceedings of the 16th International Conference on World Wide Web, Banff, Alberta, Canada, ACM, 2007, pp. 481–490. http://dx.doi.org/10.1145/1242572.1242638.

[25] Q. Mei, D. Cai, D. Zhang, C. Zhai, Topic modeling with network regularization, in: Proceedings of the 17th International Conference on World Wide Web, Beijing, China, ACM, 2008, pp. 101–110. http://dx.doi.org/10.1145/1367497.1367512.

[26] A. Micarelli, F. Sciarrone, Anatomy and empirical evaluation of an adaptive web-based information filtering system, User Model. User-Adapted Interact. 14 (2-3) (2004) 159–200. http://dx.doi.org/10.1023/b:user.0000028981.43614.94.

[27] R. Mihalcea, P. Tarau, TextRank: Bringing Order into Texts, in: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 404–411.

[28] D. Pal, M. Mitra, K. Datta, Improving query expansion using WordNet, J. Assoc. Inf. Sci. Technol. 65 (12) (2014) 2469–2478. http://dx.doi.org/10.1002/asi.23143.

[29] J.M. Ponte, W.B. Croft, A language modeling approach to information retrieval, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1998, pp. 275–281.

[30] M. Sah, V. Wade, Personalized concept-based search on the Linked Open Data, J. Web Sem. 36 (/) (2016) 32–57. http://dx.doi.org/10.1016/j.websem.2015.11.004.

[31] M. Šajgalík, M. Barla, M. Bieliková, Efficient Representation of the Lifelong Web Browsing User Characteristics, Shlomo Berkovsky and Santos [63], 2013.

[32] X. Shen, B. Tan, C. Zhai, Implicit user modeling for personalized search, in: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, Bremen, Germany, ACM, 2005, pp. 824–831. http://dx.doi.org/10.1145/1099554.1099747.

[33] M. Speretta, S. Gauch, Personalized search based on user search histories, in: Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on, 2005, pp. 622–628. http://dx.doi.org/10.1109/wi.2005.114.

[34] B. Steichen, M.R. Ghorab, A. O'connor, S. Lawless, V. Wade, Towards personalized multilingual information access – exploring the browsing and search behavior of multilingual users, in: Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization, Aalborg, Denmark, Springer, 2014, pp. 435–446. http://dx.doi.org/10.1007/978-3-319-08786-3_39.

[35] J.-T. Sun, H.-J. Zeng, H. Liu, Y. Lu, Z. Chen, CubeSVD: a novel approach to personalized web search, in: Proceedings of the 14th International Conference on World Wide Web, Chiba, Japan, ACM, 2005, pp. 382–390. http://dx.doi.org/10.1145/1060745.1060803.

[36] F. Ture, J. Lin, Exploiting representations from statistical machine translation for cross-language information retrieval, ACM Trans. Inf. Syst. 32 (4) (2014) 1–32. http://dx.doi.org/10.1145/2644807.

[37] D. Vallet, I. Cantador, J.M. Jose, Personalizing web search with folksonomy-based user and document profiles, in: Advances in Information Retrieval, Springer, 2010, pp. 420–431.

[38] E. Vicente-López, L. De Campos, J. Fernández-Luna, J. Huete, A. Tagua-Jiménez, C. Tur-Vigil, An automatic methodology to evaluate personalized information retrieval systems, User Model. User-Adapted Interact. 25 (1) (2015) 1–37. http://dx.doi.org/10.1007/s11257-014-9148-9.

[39] C.C. Vogt, G.W. Cottrell, Fusion via a linear combination of scores, Inf. Retr. 3 (October 01) (1999) 151–173. 1 http://dx.doi.org/10.1023/a:1009980820262 .

[40] Q. Wang, H. Jin, Exploring online social activities for adaptive search personalization, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM, 2010, pp. 999–1008.

[41] X. Wei, W.B. Croft, LDA-based document models for ad-hoc retrieval, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, ACM, 2006, pp. 178–185. http://dx.doi.org/10.1145/1148170.1148204.

[42] H. Xie, X. Li, T. Wang, L. Chen, K. Li, F.L. Wang, Y. Cai, Q. Li, H. Min, Personalized search for social media via dominating verbal context, Neurocomputing 172 (1/8/) (2016) 27–37. http://dx.doi.org/http://dx.doi.org/10.1016/j.neucom.2014.12.109.

[43] H. Xie, X. Li, T. Wang, R.Y.K. Lau, T.-L. Wong, L. Chen, F.L. Wang, Q. Li, Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy, Inf. Process. Manage. 52 (1) (2016) 61–72. (1//)http://dx.doi.org/http://dx.doi.org/10.1016/j.ipm.2015.03.001.

[44] S. Xu, S. Bao, B. Fei, Z. Su, Y. Yu, Exploring folksonomy for personalized search, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, Singapore, ACM, 2008, pp. 155–162. http://dx.doi.org/10.1145/1390334.1390363.

[45] C. Zhai, J. Lafferty, A study of smoothing methods for language models applied to ad hoc information retrieval, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2001, pp. 334–342.

[46] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, Adv. Neural Inf. Process. Syst. 16 (16) (2004) 321–328.

[47] D. Zhou, S. Lawless, V. Wade, Improving search via personalized query expansion using social media, Inf. Retr. 15 (3-4) (2012) 218–242. http://dx.doi.org/10.1007/s10791-012-9191-2.

[48] D. Zhou, M. Truran, T. Brailsford, H. Ashman, A hybrid technique for English-Chinese cross language information retrieval, ACM Trans. Asian Lang. Inf. Process. 7 (2) (2008) 1–35.

[49] D. Zhou, V. Wade, Latent document re-ranking, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3, Association for Computational Linguistics, 2009, pp. 1571–1580.

[50] X. Zhu, Z. Ghahramani, Learning from Labeled and Unlabeled Data with Label Propagation, Citeseer, 2002.