Query Expansion with Enriched User Profiles for Personalized Search Utilizing Folksonomy Data

Dong Zhou, Xuan Wu, Wenyu Zhao, Séamus Lawless, and Jianxun Liu

Abstract—Query expansion has been widely adopted in Web search as a way of tackling the ambiguity of queries. Personalized search utilizing folksonomy data has demonstrated an extreme vocabulary mismatch problem that requires even more effective query expansion methods. Co-occurrence statistics, tag-tag relationships, and semantic matching approaches are among those favored by previous research. However, user profiles which only contain a user's past annotation information may not be enough to support the selection of expansion terms, especially for users with limited previous activity with the system. We propose a novel model to construct enriched user profiles with the help of an external corpus for personalized query expansion. Our model integrates the current state-of-the-art text representation learning framework, known as word embeddings, with topic models in two groups of pseudo-aligned documents. Based on user profiles, we build two novel query expansion techniques. These two techniques are based on topical weights-enhanced word embeddings, and the topical relevance between the query and the terms inside a user profile, respectively. The results of an in-depth experimental evaluation, performed on two real-world datasets using different external corpora, show that our approach outperforms traditional techniques, including existing non-personalized and personalized query expansion methods.

Index Terms—Personalization, information search and retrieval, query formulation, user profiles and alert services

1 Introduction

Over the past number of years personalized search algorithms which utilize folksonomy data have attracted significant attention in the literature [1], [2]. This is partially due to the relative unavailability of users' search and click-through history to independent researchers not employed by, or engaged with, a commercial search engine. Another reason for utilizing folksonomy data is that tags are highly ambiguous, representing a typical real-world Web search scenario of short queries formulated by users. "Folksonomy" is a term typically used to describe the social classification phenomenon. Online folksonomy services are used by millions of users world-wide, enabling users to save and organize their online bookmarks with freely chosen short text descriptors. Fig. 1 shows an example page with tags and linked documents extracted from the famous *Bibsonomy*[1] website.

In current collaborative and social platforms, users can often play an active role in generating content and annotating resources through tags that collectively compose the folksonomy [3]. However, this uncontrolled tagging behaviour results in the use of an unrestricted vocabulary, which makes the search process prone to errors and omissions. In such circumstances, personalized query expansion (QE) has been widely adopted to overcome this limitation [4], [5], [6]. As pointed out by Xu et al. [2], social tags are equivalent to keywords which describe the web page in question and can be used as a substitute for a query to find that page, therefore the QE task here is, to some extent, similar to Web search QE.

Personalized QE attempts to expand the original query (in folksonomies, when simulating user searches, tags are normally used as queries) with other terms/words from a user profile that help to best represent the user's actual intent, or produce a query that is more likely to retrieve relevant documents. In personalized search utilizing folksonomy data, researchers frequently consider different term relationships, including co-occurrence statistics [5], [7], tag-tag relationships [4], [8] or the semantic relatedness of two terms [6]. In all of the above approaches, a user profile is usually needed to represent the user's interests in an individualised manner. In this context, the information stored in the user profile is typically past annotation information such as tags and annotations from social bookmarking systems. The advantage of exploiting this type of information is that it enables personalized search systems to gain rich knowledge about their users' interests and preferences due to the wealth of information that is available on social websites. In addition, as much of the information shared on social websites is public then the use of this public content should not pose a threat to users' privacy.

1. http://www.bibsonomy.org

- *D. Zhou, X. Wu, W. Zhao, and J. Liu are with the Key Laboratory of Knowledge Processing and Networked Manufacturing & School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, Hunan 411201, China. E-mail: dongzhou1979@hotmail.com, {1254909215, 719727262}@qq.com, ljx529@gmail.com.*
- *S. Lawless is with the ADAPT Research Centre, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland. E-mail: seamus.lawless@scss.tcd.ie.*

Manuscript received 19 July 2016; revised 17 Dec. 2016; accepted 4 Feb. 2017. Date of publication 13 Feb. 2017; date of current version 1 June 2017. Recommended for acceptance by Y. Chang.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2017.2668419

1041-4347 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

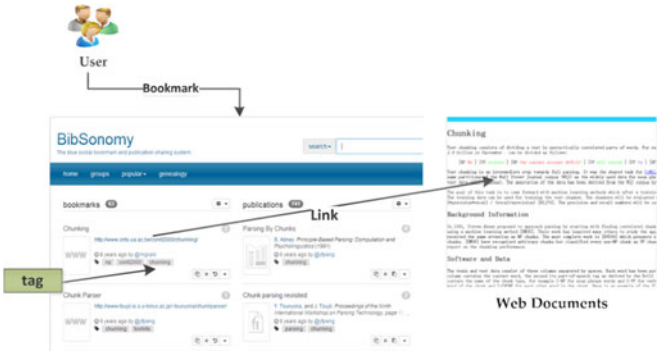

Fig. 1. Example folksonomy data.

However, some limitations exist in current approaches to personalized QE utilizing folksonomy data, including the following:

i. User profiles which contain only a user's past annotation information may not be enough to support the effective selection of expansion terms, especially for users who have had limited previous activity with the system. In this case, search personalization can be performed on an aggregate level [9]. This type of personalization involves the exploitation of usage information in a collective manner where the search process is adapted to the needs of the many, rather than the specific needs of the individual. This may "inject" the personality of other users instead of the current user, causing problems like query shift and/or interest shift. In this case, it is better to enrich the user profile according to the specific needs of the particular user rather than borrow information from similar counterparts.

ii. Previous personalized QE research either favors tag-tag relationships or relies on the co-occurrence statistics of two terms. Given the fact that tags may not constitute precise descriptions of resources, and that methods based on pure lexical matching may miss important semantic information, the retrieval performance is generally unsatisfactory [4], [5], [7], [8]. There have been a few attempts made to use topic models to capture the semantic relationships of the terms [6]. However, as pointed out by other researchers [10], [11], the observation of semantic coherence found in the inferred topic distributions by topic models is somewhat accidental. It tends to describe the statistical relationship of occurrences rather than real semantic information embedded in words. Instead, in approaches like word embeddings (WEs), words with similar syntactic and semantic properties are found to be close to each other in the embedding space. Despite the usefulness of this approach, WEs ignores the complicated correlations among words. Therefore, there is a desire to integrate the two models together to produce better relationships between terms and/or words.

In this paper, we adopt a different approach to personalized QE utilizing folksonomy data. In our approach, the expansion process is based on an enriched user profile, which contains tags and annotations together with documents retrieved from an external corpus. This corpus can be viewed as a knowledge base to enhance the information stored in the user profile. The whole procedure of query adaptation is hidden to the user. It happens in an implicit way based on their choices of tags and the terms used on annotated web pages. We first propose a novel model to build the enriched user profiles. Our model integrates the current state-of-the-art text representation learning framework, known as word embeddings, with topic models in two groups of pseudo-aligned documents between user annotations and documents from the external corpus. We then present two novel QE techniques. The first technique approaches the problem by using topical weights-enhanced WEs to select the best possible expansion terms. The second method is based on the topics learned. It calculates the topical relevance between the query and the terms inside a user profile.

This paper describes an in-depth experimental evaluation using two different real-world folksonomy datasets, extracted from *del.ico.us*[2] and *Bibsonomy*. We also explore two different external corpora for user profile enrichment. A comparative analysis of our findings with those obtained by using well-known and state-of-the-art techniques such as those exploiting co-occurrence statistics, tag-tag relationships and semantic relatedness for personalized QE, shows that our approach is able to achieve significantly better retrieval results.

The contribution of this paper can be summarized as follows:

i. We tackle the challenge of personalized QE utilizing folksonomy data in a novel way by integrating latent and deep semantics.

ii. We propose a novel model that integrates word embeddings with topic models to construct enriched user profiles with the help of an external corpus.

iii. We suggest two novel personalized QE techniques based on topical weights-enhanced word embeddings, and the topical relevance between the query and the terms inside a user profile. The techniques demonstrate significantly better results than previously proposed non-personalized and personalized QE methods.

The rest of this paper is organized as follows. Related work on personalized search utilizing folksonomy data is summarized in Section 2. Section 3 lays out the problem definition. Section 4 presents details of the user profile construction process. Section 5 demonstrates the proposed personalized QE methods. In section 6, a report on a series of experiments performed to evaluate the personalized QE strategies is provided. Finally, Section 7 concludes the paper.

2 Related Work

Web users may not always be successful in using a representative vocabulary when locating objects in a system. Therefore, query expansion attempts to expand the terms of the user's query with other terms, with the aim of retrieving more relevant results. QE has a long standing history in Information Retrieval (IR) and web search [12]. Among the various QE approaches presented in literature, some take

2. http://www.delicious.com

advantage of implicit relevance feedback [13], some use external sources [14], and some implement semantic QE [15]. These techniques are generally non-user focused. There are also user-focused QE methods. For example, methods that implicitly select terms from the user profile [16], methods which involve implicitly obtaining terms from the query logs and/or their associated clicked documents [17], and methods requiring the user to explicitly provide relevance feedback or perform interactive query expansion [18].

Folksonomy-based systems have gained great popularity and attention in recent times. Personalized QE utilizing folksonomy data primarily considers term relationships from an individual perspective or in an aggregate manner. Researchers have considered tag-tag relationships for personalized QE, by selecting the most related tags from a user's profile [4], [8], [19]. However, tags might not be precise descriptions of web pages, and as a result the retrieval performance of this QE approach is somewhat disappointing. Local analysis and co-occurrence based user profile representation have also been adopted to expand the query according to a user's interaction with the system [5], [7]. It is worth noting that in [5], folksonomy data are not used as a test bed as in other approaches, but rather used as an external source of information from which to extract semantic classes that are added to web search results. Moreover, terms in this approach are still based on co-occurrence statistics rather than semantic relatedness. Zhou et al. proposed a personalized QE framework based on the semantic relatedness of terms inside individual user profiles [6]. A statistical tag-topic model is created to deduce latent topics from the user's tags and tagged documents. This model is then used to identify the most relevant terms in the user model to the user's query and then use those terms to expand the query.

Our approach in this paper also considers the semantic relatedness between terms inside user profiles. It differs from past research in two aspects. First, we exploit an external knowledge base to enrich the user profiles, while previous research builds them from historical usage information alone. Second, we consider recent advances in neural language models in addition to topic models for personalized QE.

The literature proposes several systems that do not perform query expansion but still use folksonomies to provide users with personalized search services. The authors in [20] investigated re-ranking results retrieved from the Yahoo search engine based on a user profile comprising tags extracted from the user's participation on the *del.icio.us* social bookmarking website. A similar approach was also explored in [21] where the system performed re-ranking of Google search results based on social bookmarks and tags harvested from *del.icio.us*. However, the data sparsity problem poses a challenge to this approach as not all Web pages returned by search engines are tagged in the *del.icio.us* dataset.

Because of this data sparsity problem, researchers started to use folksonomy data as a test collection to develop personalized techniques. In [2] the authors developed a personalization approach to learn about users' interests from their bookmarks and tags, then re-rank the results according to the topic relevance of documents and users' interests. Wang and Jin [22] explored gathering data from multiple online social systems for adaptive search personalization. Bouadjenek et al. [23], [24] propose using social data and user relationships to enhance document representation for re-ranking purposes. Cai et al. [25] model query relevance measurement and user relevance measurement as fuzzy satisfaction problems. Verbal context [26] and sentiment analysis [27] are also utilized. Note that in the last three approaches the actual content of the web pages/documents are not used, as opposed to the approaches proposed by this paper.

3 Problem Definition

Web 2.0 leverages the Web as a collaborative and social platform. In folksonomy applications like social tagging systems, users can label interesting web resources with primarily short and unstructured *annotations* in natural language called *tags*. These web resources are denoted as a URL in the *del.icio.us* or *BibSonomy* website. Textual content can be crawled by following the URL that refers to a *document* or *web page*. We present the basic notations used in this paper in Table 1.

Formally, data in folksonomy systems can be represented by a tuple $\mathcal{P} := (\mathcal{U},\ \mathcal{D},\ \mathcal{T},\ \mathcal{A})$. $\mathcal{A} \subseteq \mathcal{U} \times \mathcal{D} \times \mathcal{T}$ is a ternary relation, whose elements are called tag assignments or annotations. The set of annotations of a user is defined as: $\mathcal{A}^u := \{(t,\ d)|u,\ d, t \in \mathcal{A}\}$. The tag vocabulary of a user, is given as $\mathcal{T}^u := \{t|(t,\ d) \in \mathcal{A}^u\}$. A set of documents annotated by a user u is defined as $\mathcal{D}^u := \{d|(t,\ d) \in \mathcal{A}^u\}$. We define the terms extracted from a user's set of documents as $term^{\mathcal{D}^u} := \{w|w \in \mathcal{D}^u\}$, where w denotes a word/term in the annotated documents. Similarly, we define terms extracted from a user's set of external documents as $term^{\mathcal{D}^u_{exter}} := \{w|w \in \mathcal{D}^u_{exter}\}$, where $\mathcal{D}^u_{exter}$ denotes a user's set of external documents from an external corpus $\mathcal{D}_{exter}$.

In a typical personalized search scenario, given a source query q and a set of words in the user model $\{w_1,\ w_2 \ldots w_n\} \in term^{\mathcal{D}^u} \cup term^{\mathcal{D}^u_{exter}} \cup \mathcal{T}^u$ the goal is to return a ranked list of terms to be added to the query, for a second round retrieval of results.

4 Enriched User Profiles

The enriched user profile generation consists of two stages: external document retrieval and user profile construction. We enrich a user's historical usage information with documents retrieved from an external corpus. This procedure is described in Algorithm 1. We first concatenate all tags t in $\mathcal{T}^u$ into one single query $q^{\mathcal{T}^u}$ representing a user's past interests through his/her tags (Algorithm 1, line 1). Then for each document d^u in a user's set of documents in the user profile $\mathcal{D}^u$, we extract terms with the highest inverted document frequency (*idf*) scores[3] as a query q^{d^u} (Algorithm 1, lines 2-4, with the *extractTop* function returns top λ terms). In other words, each $q^{d^u}(d^u \in \mathcal{D}^u)$ contains representative terms from document d^u that a user has tagged. Next we send queries in Q_{exter} (contains $q^{\mathcal{T}^u}$ and all q^{d^u}) to an external corpus $\mathcal{D}_{exter}$ to fetch $\mathcal{D}^u_{exter}$ (Algorithm 1, lines 5-8, the number of documents retrieved by each query is controlled by the parameter γ).

3. The reason we use *idf* here is to assign more weight to words bearing more information content. Other alternatives exist, such as term frequency (*tf*) or *tf-idf*, however, through testing they have been deemed less effective in this context.

TABLE 1
Basic Notations Used in the Paper

Symbol	Meaning	Symbol	Meaning
$\mathcal{U}$	finite sets of users	$\vec{w}$	pivot word representation or pivot word embedding of w
$\mathcal{D}$	finite sets of web pages/documents	dim/x	dimensionality of word embeddings
$\mathcal{T}$	finite sets of tags	K	number of topics
$\mathcal{A}$	a ternary relation, elements are tags	θ	multinomial distribution of topics
$\mathcal{A}^u$	the set of annotations of a user	φ	multinomial distribution of words
$\mathcal{T}^u$	the tag vocabulary of a user	ϕ	multinomial distribution of words (another group)
$\mathcal{D}^u$	a user's set of documents	α	the parameter of topic Dirichlet prior
t	a tag	β	the parameter of word Dirichlet prior
d	a document	d_j^C	a document in group C
u	a user	d_j^G	a document in group G
w	a word/term	N_{d_j}	number of words in document d_j (group indicator omitted)
$term^{\mathcal{D}^u}$	the vocabulary extracted from the documents that a user has tagged	$z_{j,i}$	topic associated with the ith word in the document d_j
$term^{\mathcal{D}^u_{exter}}$	the full set of terms extracted from a user's external documents	$w_{j,i}$	ith word in document d_j
$\mathcal{D}_{exter}$	an external corpus	$f^e_{j,i}$	Dimension e of the embedding of word $w_{j,i}$
$\mathcal{D}^u_{exter}$	a user's set of external documents	μ_z	mean of Log-normal distribution of retrieval scores for topic z
q	a source query	σ_z	deviation of Log-normal distribution of retrieval scores for topic z
$q^{\mathcal{T}^u}$	a query containing the concate nated tags of a user	$n_{j,k}$	the number of times that topic k is sampled w.r.t. document d_j
q^{d^u}	a query extracted from a document d^u that a user tagged	$v_{k,w_{j,i}}$	the number of times $w_{j,i}$ has been generated by topic k
Q_{exter}	queries to be sent to an external corpus	R	an underlying hypothetical model of relevance
C	source group of pseudo-aligned documents	$\vec{q}$	vector representation of a query q
G	target group of pseudo-aligned documents	tw_i	weight of word w_i

Algorithm 1. External document fetch

Require: tags of a user $\mathcal{T}^u$
Require: documents of a user $\mathcal{D}^u$
Require: an external corpus $\mathcal{D}_{exter}$
1. $q^{\mathcal{T}^u} \leftarrow \bigcup (t \in \mathcal{T}^u)$
2. $Q_{exter} \leftarrow q^{\mathcal{T}^u}$
3. **for each** $d^u \in \mathcal{D}^u$ **do**
4. $\quad q^{d^u} \leftarrow extractTop(w \in d^u)$
5. $\quad Q_{exter} \leftarrow q^{d^u}$
6. **end for**
7. **for all** $q \in Q_{exter}$ **do**
8. $\quad \mathcal{D}^u_{exter} \leftarrow retrieve_{\mathcal{D}_{exter}}(q)$.
9. **end for**

In stage two, we integrate $\mathcal{T}^u$ (here all tags are concatenated and viewed as a single document), $\mathcal{D}^u$ and $\mathcal{D}^u_{exter}$ into a novel generative model such that a multinomial distribution of topics specific to each document can be inferred. We now describe this procedure in detail.

It is well known that the Latent Dirichlet Allocation (LDA) model and its extensions play an important role in natural language processing and machine learning by mining the thematic structure of documents [28]. However, the probability distribution from LDA only describes the statistical relationship of occurrences in the corpus. Recently, word embeddings have begun to play an increasingly vital role in building continuous word vectors based on their contexts in a corpus. It has been shown that in some applications, the embedded representations are more effective than representations produced by the LDA model [29]. There are also some attempts to integrate LDA with WEs for different purposes [10], [11]. Inspired by those works as well as work in bilingual documents [30], we propose a novel generative model for user profile generation based on the documents obtained in the last stage. We named this *enriched user profile construction* (EUPC) model.

In this model, $\mathcal{T}^u$, $\mathcal{D}^u$ and $\mathcal{D}^u_{exter}$ can be mixed together to infer unified latent topics. However, as $\mathcal{D}^u_{exter}$ is actually different from the user's original information $\mathcal{T}^u$ and $\mathcal{D}^u$, it is better to model them separately rather than jointly. Moreover, the model should be able to learn latent topics from non-parallel data. With this consideration in mind, EUPC learns topics which are shared between document-aligned pairs. In order to do this, we create pseudo-aligned documents between $\mathcal{T}^u$, $\mathcal{D}^u$ and $\mathcal{D}^u_{exter}$. This procedure works as follows. For each external document in $\mathcal{D}^u_{exter}$ retrieved by a query from Q_{exter}, which is formed through stage one of our approach, we treat the document (from $\mathcal{D}^u_{exter}$) and the documents that generate Q_{exter} (from $\mathcal{T}^u$ and $\mathcal{D}^u$) as pseudo-aligned documents in two groups. The first group we named source group C, the other group we named target group G. By using the aligned documents, we propose a model to learn the latent topics between the two groups.

To jointly model words and word embeddings, EUPC learns a shared latent topic space to generate words in

documents and corresponding word embeddings. The model takes pre-trained word embeddings and documents as input. In other words, embeddings are given as observed variables in our model. We use the Skip-Gram model [31] to learn the WEs before running our model. Skip-gram aims to predict context words given a target word in a sliding window. Each target word w is associated with a vector $\vec{w} \in \mathbb{R}^{dim}$ (its pivot word representation or pivot word embedding). dim is the dimensionality of WEs. The vector of the target word is used as a feature to predict the context words. We employ a normal distribution for WEs to infer latent topics via the documents and words.

Algorithm 2. Generative process for EUPC

Require: tags of a user $\mathcal{T}^u$
Require: documents of a user $\mathcal{D}^u$
Require: a user's set of external documents $\mathcal{D}^u_{exter}$
Require: word embeddings calculated by Skip-Gram for all words in $\mathcal{T}^u \cup \mathcal{D}^u \cup \mathcal{D}^u_{exter}$

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. **end for**
5. **for** each document pair $d_j = \{d_j^C \in \mathcal{T}^u \cup \mathcal{D}^u, \quad d_j^G \in \cup \mathcal{D}^u_{exter}\}$ **do**
6. sample the mixture of topics $\theta_j \sim Dirichlet(\alpha)$
7. **for** each word w_i^C indexed by $i = 1, \ldots, N_{d_j^C}$ **do**
8. sample the topic index topic $z_{j,i}^C \sim Mult(\theta_{d_j})$
9. sample the term for word $w_{j,i}^C \sim Mult(\varphi_{z_{j,i}^C})$
10. for each dimension of the embedding of $w_{j,i}^C$, sample $f_{j,i}^{e^C} \sim \mathcal{N}(\mu_{z_{j,i}}^{e^C}, \sigma_{z_{j,i}}^{e^C})$
11. **end for**
12. **for** each word w_i^G indexed by $i = 1, \ldots, N_{d_j^G}$ **do**
13. sample the topic index topic $z_{j,i}^G \sim Mult(\theta_{d_j})$
14. sample the term for word $w_{j,i}^G \sim Mult(\phi_{z_{j,i}^G})$
15. for each dimension of the embedding of $w_{j,i}^G$, sample $f_{j,i}^{e^G} \sim \mathcal{N}(\mu_{z_{j,i}}^{e^G}, \sigma_{z_{j,i}}^{e^G})$
16. **end for**
17. **end for**

With the documents and WEs trained by the Skip-Gram model, the generation process of the EUPC model can be summarized as follows (see Algorithm 2). In this case, there is a comparable document set aligned at the document-level. θ can be viewed as a group independent factor. First, the model generates a mixture of words into the different group of documents (Algorithm 2, lines 1-4 Then for each document pair d^C and d^G, a shared topic distribution is chosen from a Dirichlet distribution (Algorithm 2, line 5-6). Next, the word distributions are chosen for each of the topics selected in the previous step for both documents. In each document in group C, for each word i in document d_j^C, a particular topic $z_{j,i}^C$ can be sampled from the document-specific distribution, a word indicator $w_{j,i}^C$ is drawn from the topic-specific distribution $\varphi_{z_{j,i}}^C$ and for each dimension e of the embedding of word $w_{j,i}^C$, we draw $f_{j,i}^{e^C}$ from normal distribution $\mathcal{N}(\mu_{z_{j,i}}^{e^C}, \sigma_{z_{j,i}}^{e^C})$ (Algorithm 2, lines 7-11). The process is repeated for all words in the document and the procedure is repeated for the document in group G also.

In the Algorithm 2, μ and σ are the mean and deviation of the normal distribution. α and β are the parameters of topic Dirichlet prior and word Dirichlet prior. θ_j is the multinomial topic distribution of document pair d^C and d^G. We used a fixed number of latent topics and dimensions for WEs. The posterior distribution of topics depends on two sets of information, both the words and WEs.

In this model, inference is intractable. We use Gibbs Sampling to perform approximate inference. We employ a conjugate prior for the multinomial distributions, and integrate out θ, φ and ϕ. In the sampling procedure, we need to calculate the conditional distribution $p(\mathrm{z}_{j,i}^C = k)$ and $p(\mathrm{z}_{j,i}^G = k)$. By using Gibbs Sampling, for each word the topic is sampled from:

$$
\begin{aligned}
&p(z_{j,i}^C = k) \\
&\propto \frac{n_{j,k,\neg i}^C + n_{j,k}^G + \alpha}{n_{j,\cdot,\neg i}^C + n_{j,\cdot}^G + K \cdot \alpha} \times \frac{v_{k,w_{j,i}^C,\neg}^C + \beta}{v_{k,\cdot,\neg}^C + V^C \cdot \beta} \\
&\times \prod_{e=1}^{dim} \frac{1}{\sqrt{2\pi}\sigma_{z_{j,i}}^C} \exp\left(- \frac{\left(f_{j,i}^{e^C} - \mu_{z_{j,i}}^C\right)^2}{2{\sigma_{z_{j,i}}^C}^2} \right)
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
&p(z_{j,i}^G = k) \\
&\propto \frac{n_{j,k,\neg i}^G + n_{j,k}^C + \alpha}{n_{j,\cdot,\neg i}^G + n_{j,\cdot}^C + K \cdot \alpha} \times \frac{v_{k,w_{j,i}^G,\neg}^G + \beta}{v_{k,\cdot,\neg}^G + V^G \cdot \beta} \\
&\times \prod_{e=1}^{dim} \frac{1}{\sqrt{2\pi}\sigma_{z_{j,i}}^G} \exp\left(- \frac{\left(f_{j,i}^{e^G} - \mu_{z_{j,i}}^G\right)^2}{2{\sigma_{z_{j,i}}^G}^2} \right)
\end{aligned} \tag{2}
$$

for each group of documents, $\mathrm{n_{j,k,\neg i}}$ counts the number of times that a topic with index k has been sampled from the multinomial distribution specific to document $\mathrm{d_j}$ with the current $\mathrm{z_{j,i}}$ not counted. Another counter variable $\mathrm{v_{k,w_{j,i},\neg}}$ counts the number of times $\mathrm{w_{j,i}}$ has been generated by topic k, but not counting the current $\mathrm{w_{j,i}}$. A dot denotes summation over all values of the variable whose index that dot takes. After that we can calculate the posterior estimate of θ, φ and ϕ. Fig. 2 shows a graphical representation of the EUPC model.

5 Personalized Query Expansion

In this section, we present our new personalized QE techniques. One technique is based on the posterior estimation of word-topic distributions and WEs generated, while another technique is based on the topics learned by using the EUPC model. The first method uses only the EUPC model (integration of topic models with WEs) to weight the word representations produced by WEs. The second method, however, fully exploits the advantages of the integration of the two semantic models.

5.1 Query Expansion Based on Word Embeddings

Our first QE approach, Query Expansion based on Word Embeddings (WEQE), works as follows. In order to present the query q in the x-dimensional embedding space induced

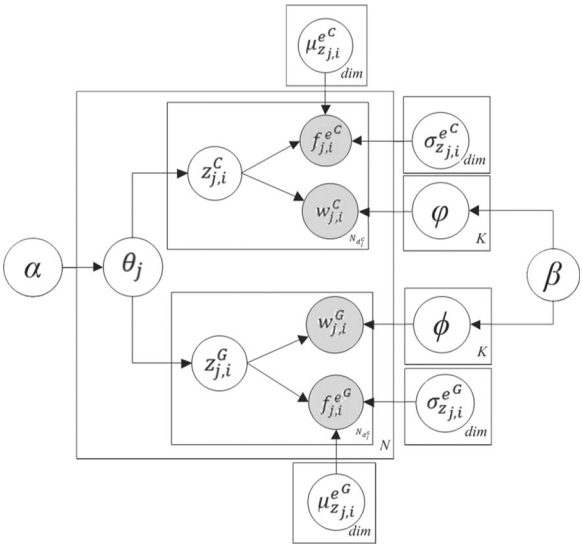

Fig. 2. Plate notation of the EUPC model.

$(x = dim)$, we apply a model of semantic composition to learn its x-dimensional vector representation $\vec{q}$. We opt for a simple addition as the compositional operator [33]. As before, assume the query q consists of n independent query terms:

$$\vec{q} = \overrightarrow{w_1} + \overrightarrow{w_2} + \cdots + \overrightarrow{w_n} \tag{3}$$

By representing the query in the x-dimensional embedding space, the similarity between the words inside a user profile and the query can be easily calculated. In this paper, we used the standard cosine similarity measure, the standard choice in the embedding induction literature, as a similarity function [29], [31] as follows[4]:

$$sim(\vec{w}, \vec{q}) = \cos(\vec{w}, \vec{q}) = \frac{\vec{w} \cdot \vec{q}}{|\vec{w}| \cdot |\vec{q}|} \tag{4}$$

However, this calculation does not consider the output of the EUPC model. In an extended approach, we weight the word vectors with the posterior estimation of word-topic distributions. Take a word $w_{j,i}^C$ from group of C of the EUPC model as an example, we obtain the weight of the word as follows: $tw_i^C = \sum_{k=1}^{K} \varphi_{i,k}$. We compare this weighting scheme with other simpler alternatives in Section 6.2.4. A word embedding is then constructed out of its WEs as in Equation (5):

$$\overrightarrow{w_i} = tw_i^C \cdot \overrightarrow{w_i} \tag{5}$$

All the profile terms $\{w_1,\ w_2 \dots w_n\}$ are ranked by their similarities to the given query, and the top δ terms are chosen to expand the query. The procedure is summarized in Algorithm 3.

5.2 Topical Query Expansion

In the topical query expansion (TQE) approach, we calculate the weights of the terms from a user profile to be added to the initial query. The key idea is that both documents containing those profile terms and query terms are assumed to be sampled from an underlying hypothetical model of relevance R (see also [13], [32]).

Algorithm 3. Query Expansion based on Word Embeddings

Require: a user's set of documents $term^{\mathcal{D}^u}$
Require: a user's set of external documents $term^{\mathcal{D}^u_{exter}}$
Require: the tag vocabulary of a user $\mathcal{T}^u$
1. $\{w_1,\ w_2 \dots w_n\} \leftarrow term^{\mathcal{D}^u} \cup term^{\mathcal{D}^u_{exter}} \cup \mathcal{T}^u$
2. **for all** $w \in \{w_1,\ w_2 \dots w_n\}$ **do**
3. calculate $sim(\vec{w}, \vec{q})$
4. **end for**
5. **return** q' consists of top δ terms with the highest $sim(\vec{w}, \vec{q})$

Given the query $q = \{w_a\}_{a=1}^n$ of n independent query terms, the probability of the query generating a word w from an underlying model R is approximated by:

$$P(w|R) \approx P(w|q) \tag{6}$$

We further assume that there are a set of relevant documents $\{d_b\}_{b=1}^N$ related to the query and the word being considered, where N is the number of documents. Incorporating this set of documents into the above equation leads to,

$$\begin{aligned} P(w|q) &= \sum_{b=1}^{N} P(w|d_b, q) P(d_b|q) \\ &= \sum_{b=1}^{N} P(w|d_b) P(d_b|q) \\ &\propto \frac{1}{N} \sum_{b=1}^{N} P(w|d_b) P(q|d_b) \\ &= \frac{1}{N} \sum_{b=1}^{N} P(w|d_b) \prod_{a=1}^{n} P(w_a|d_b). \end{aligned} \tag{7}$$

The calculation discards the uniform prior for $P(q)$, and takes the uniform prior of documents $P(d_b) = \frac{1}{N}$ outside the summation. It assumes that query terms are sampled independently and identically to each other. It also assmues that $P(w|d_b, q) = P(w|d_b)$. Note that this latter independent assumption applies to all the remaining equations in this section.

As we already have outputs from the EUPC model, the documents inside the user profile can be used as a set of relevant documents in the above calculation. The Equation (7) has an intuitive explanation in the sense that the likelihood of generating a word w from the document model will increase if the numerator $P(w|d_b) \prod_{a=1}^{n} P(w_a|d_b)$ increases, or in other words if w co-occurs frequently with the query terms in the relevant document d_b. This model thus utilizes the co-occurrence of a non-query term with the given query to boost the retrieval scores of documents. If this model is not utilized here, we will get a lower language model similarity score due to vocabulary mismatch between non-query terms and query terms.

However, the above calculation has an oversimplified assumption that each relevant document is generated from a single generative model. A query typically encompasses multiple aspects of the overall information need expressed

4. Clearly other options, such as summing or averaging the elements of the word representations, could be used. These will be explored in future work.

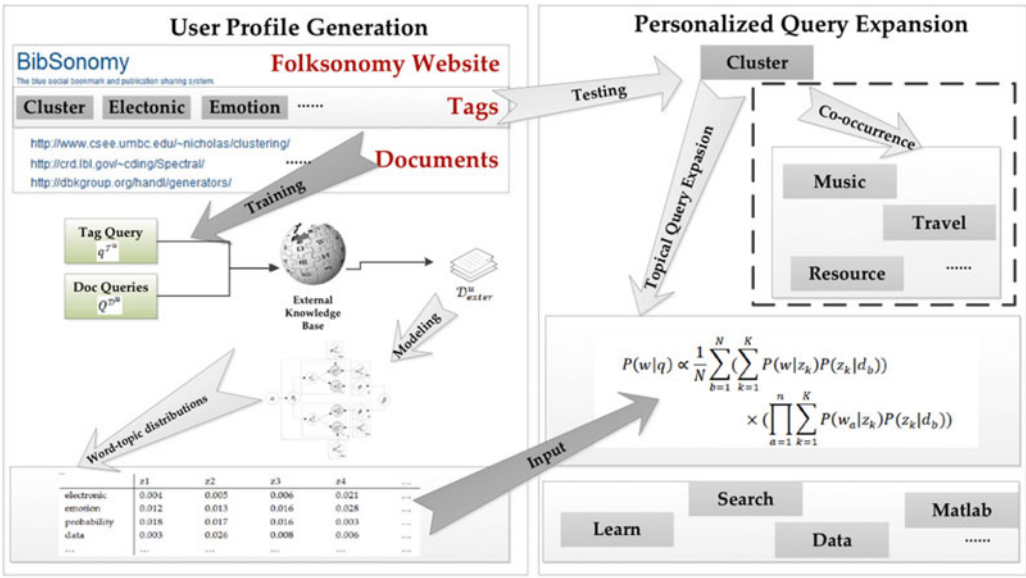

Fig. 3. Overview of the QE process.

in it. Thus in a more general case, it would be reasonable to assume that the query terms are sampled from a number of relevance models instead of one.

We now have latent topics related to each document and each word. Therefore there is no longer a direct dependency of w and w_a on d_b. In this case, in order to estimate $P(w|d_b)$, we can marginalize the probability over the latent topic variables z_k. Then we have,

$$P(w|d_b) = \sum_{k=1}^{K} P(w|z_k)P(z_k|d_b). \tag{8}$$

Similarly, the probability $P(w_a|d_b)$ becomes,

$$P(w_a|d_b) = \sum_{k=1}^{K} P(w_a|z_k)P(z_k|d_b). \tag{9}$$

So that the probability of the query generating a word w can be re-defined as,

$$P(w|q) \propto \frac{1}{N}\sum_{b=1}^{N}\left(\sum_{k=1}^{K} P(w|z_k)P(z_k|d_b)\right) \times \left(\prod_{a=1}^{n}\sum_{k=1}^{K} P(w_a|z_k)P(z_k|d_b)\right). \tag{10}$$

This calculation thus involves two levels of EUPC model estimated term generation probabilities, one for the words in relevant documents and the other for the query terms. It ensures that it assigns higher probability to a term being generated from the document model, if the term co-occurs with query terms in relevant documents and is also likely to belong to the same topic as that of the query terms.

In our EUPC model, we use one side of the word-topic distributions from the group that contains tags and annotated documents to calculate the weighting. All the profile terms $\{w_1,\ w_2 \dots w_n\}$ are ranked by their probability of being generated by the given query, and the top δ terms are chosen to expand the query. The procedure is summarized in Algorithm 4.

Algorithm 4. Topical Query Expansion

Require: the vocabulary extracted from the documents that a user has tagged $term^{\mathcal{D}^u}$
Require: the full set of terms extracted from a user's external documents $term^{\mathcal{D}^u_{exter}}$
Require: the tag vocabulary of a user $\mathcal{T}^u$
1. $\{w_1,\ w_2 \dots w_n\} \leftarrow term^{\mathcal{D}^u} \cup term^{\mathcal{D}^u_{exter}} \cup \mathcal{T}^u$
2. **for all** $w \in \{w_1,\ w_2 \dots w_n\}$ **do**
3. calculate $P(w|q)$
4. **end for**
5. **return** q' consists of top δ terms with the highest $P(w|q)$

To give an overview of our proposed QE process, we plot a figure consisting of our user profile generation and personalized query expansion methods in action in Fig. 3 (for simplicity we only show topical query expansion in the figure). We also provide a very simple example in the figure. As we can see, for query "cluster", a co-occurrence-based method (included in the dashed box for comparison, not part of our approach, see Section 6) without user profile enrichment can only produce expansion terms according to the bookmark statistics like "music", "travel", etc. (note only representative terms shown). However, our methods can generate expansion terms that are semantically related to the query and better reflect the user's true personal needs such as "data", "matlab", etc. By checking the actual web resources the particular user tagged, we found that indeed she was looking for effective methods to do data clustering. In the next section we will systematically evaluate our methods and compare to several state-of-the-art baseline models.

6 Evaluation

6.1 Experimental Setup

In order to evaluate the proposed methods above on real-world data, we selected two real-world folksonomy datasets from *del.icio.us* and *BibSonomy*. The first test collection (referred to as *DEL*) was constructed from *del.icio.us* using the following two sub datasets: *socialbm0311* and *deliciousT140*, which are public and described and analyzed in

TABLE 2
Statistics for Two Test Collections

	Users	Documents	Tags
del.icio.us	259,511	131,283	137,870
BibSonomy	27,611	143,686	59,799

TABLE 3
Statistics for External Corpora

	Documents	Terms
Wikipedia	4,634,369	5,357,496
CLEF News	304,630	406,244

[34], [35]. The *deliciousT140* dataset contains 144,574 unique URLs, all of them with their corresponding social tags retrieved. However, this dataset does not contain the actual web pages. So we used another dataset, *socialbm0311*, which is a large-scale social tagging/bookmarking dataset which contains the complete bookmarking activity for almost 2 million users. After matching the documents in *deliciousT140 wi*th the bookmark activities in *socialbm0311*, we obtained a total of 5,153,720 bookmark activities, 259,511 users, 131,283 web pages and 137,870 tags. The second test collection is constructed from the latest *BibSonomy* dumps on 01/01/2016 (referred to as *BIB*). It contains 3,887,070 bookmark activities. Similar to the *socialbm0311*dataset, the *BibSonomy* dataset does not contain the actual web pages. So we crawled the Web for those web pages that were still available, finally obtaining 27,611 users, 143,686 web pages and 59,799 tags. We used a public parser[5] to parse all the web pages in order to get their textual content. A statistical summary of both test collections is presented in Table 2.

We also constructed two corpora from different external knowledge bases. The first corpus was obtained from the largest online encyclopedia – Wikipedia[6]. A Wikipedia snapshot was obtained on the 14/08/2014, which contained a collection of 4,634,369 articles. The second corpus consists of English news documents from the Glasgow Herald 1995, Los Angeles Times 1994 and Los Angeles Times 2002, a collection made available by the CLEF AdHoc-News Test Suites (2004-2008)[7], which we refer to as CLEF News. We present a statistical summary of both external corpora in Table 3.

We selected three groups of users as test users: users with no more than 50 bookmarks (referred to as *U50*), users with 50-500 bookmarks (referred to as *U500*), and users with more than 500 bookmarks (referred to as *UG500*). These groups of users represent users with small, moderate and rich amounts of past usage information, respectively. 200 randomly selected users from each group are chosen as test users. For each user, 75 percent of his/her tags with annotated web pages were used to create the user profile and the other 25 percent were used as a test collection. Another subset of users was also randomly selected from the dataset to train the necessary parameters. Every effort was made to ensure there was no overlap between the group of users used for parameter training and the three groups of test users.

The evaluation method used by previous researchers in personalized social search [2], [6], [22] is employed. The main assumption is as follows: Any documents tagged by u with t are considered relevant for the personalized query (u, t) (u submits the query t). In particular, we follow the same procedure as in [2], [6], [22] to use the annotations to generate queries. As pointed out by [2], in today's search technology, keyword queries are the most popular query paradigm, and social tags are generally good keyword descriptors of the web pages in question. Moreover, tags reflect the users' personal preferences with regard to vocabulary, often the vocabulary they use in daily life. Hence the data is not biased toward the experiments performed in this paper.

The following evaluation metrics were chosen to measure the effectiveness of the various approaches: normalized discounted cumulative gain (NDCG, we report NDCG@10 results), mean reciprocal rank (MRR) and mean average precision (MAP). The average performance over all users is calculated. Statistically significant differences were determined using a paired t-test at a confidence level of 95 percent.

We evaluate our proposed models and compare with several state-of-the-art non-personalized and personalized query expansion methods as follows.

LanM. A popular and quite robust language model retrieval method, which has previously demonstrated good results. We compute the Kullback-Leibler divergence between the query language model and document language model as described in [36].

RelM. A relevance model which involves pseudo-relevance feedback in the language model as in [13]. We include this model as a competitive non-personalized query expansion baseline.

ExtRelM. This is a modified version of the relevance model described in [14]. Instead of using the top-ranked documents as pseudo-relevance documents, this model uses external corpora to obtain the relevance documents. We include this model as a strong non-personalized baseline as we also used external corpora in our models. In the experiments, this method will acquire external documents from the Wikipedia and CLEF News corpora.

CoWM. This method has been used by several re searchers [5], [7]. In this method the selection of expansion terms is based on co-occurrence statistics between the query terms and other terms inside the user model. We used this approach as it previously demonstrated satisfactory performance [16].

CoTagM. Pure tag-tag relationships are also favored by many researchers. This method is based on the co-tagging activities a user performed [4], [8], [19]. In this case, the user profiles contain training tags with their co-tagging statistics computed using the Jaccard coefficient as in [19].

TagTM. Zhou et al. [6] proposed a query expansion framework based on semantic word associations enhanced by using terms extracted from top-ranked documents. The user profiles are built according to a Tag-Topic model for all profile terms. We include the highest performing method from their work for comparison.

EnUWEM. From our proposed methods, the first method uses the EUPC model and the WEQE method to personalize search.

5. http://htmlparser.sourceforge.net/
6. http://www.wikipedia.org
7. http://catalog.elra.info/product_info.php?products_id=1127

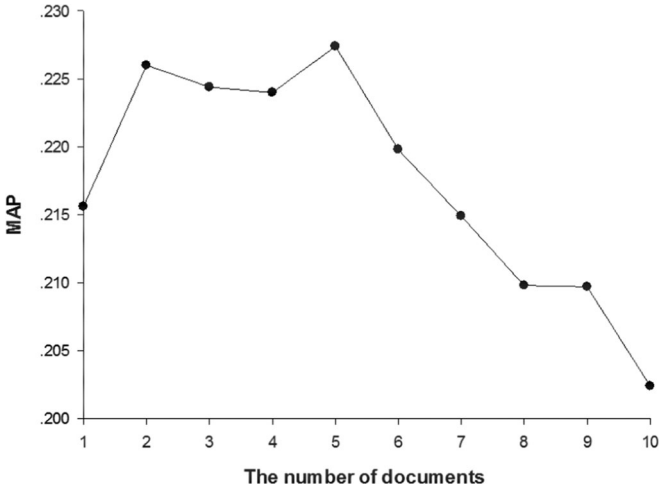

Fig. 4. Tuning parameter γ.

EnUTM. This is our alternative proposed method, which uses the EUPC model and the TQE method for personalized search utilizing folksonomy data.

As the number of documents retrieved by each query in step one (i.e., γ) is deemed to be an important parameter, we test in the range [1], [10]. It can be seen from Fig. 4 that measured by MAP there are no significant differences when the number is between 2 and 5. There is, however, a noticeable drop when the number of documents goes beyond 5, which means too many documents may introduce noise. We simply choose the number that shows the best performance, which sets γ to 5. Parameter λ used in the *extractTop* function is set to 10. For the EUPC modeling, α and β were set to $50/K$ and 0.01. In the expansion method, the number of expansion terms δ are set to 50. The number of word dimensions and latent topics used in *EnUWEM* and *EnUTM* are set to 80 and 20 empirically for all test collections. All the parameters in the other baseline models are set according to their tuning procedures in the original papers or those obtaining best performance.

6.2 Results and Discussion

6.2.1 Performance in the DEL Collection, Using Wikipedia as External Corpus

In this section, we first report the results obtained by using the *DEL* collection and then the *BIB* collection. The experimental results that describe the performance of the proposed methods in this paper together with three non-personalized baselines on the overall test users in the *DEL* collection within different groups and Wikipedia as external corpus are shown in Table 4. The statistically significant differences are marked as *r* and *e* with respect to the *RelM* and *ExtRelM* baselines as these two methods work better than the simpler *LanM* method. Clearly all personalized approaches performed significantly better than non-personalized methods, including our proposed approaches *EnUWEM* and *EnUTM*. This illustrates that non-personalized query expansion methods can only bring limited improvements, especially for search utilizing folksonomy data as tags introduce more ambiguity. The methods which use additional terms from user profiles can greatly improve retrieval effectiveness.

Next we evaluate the performance of the proposed methods compared to several personalized baselines, i.e., the *CoWM*, *CoTagM*, and *TagTM* methods. The statistically significant differences in the table are marked as *w*, *c*, and *t* respectively.

TABLE 4
Overall Results on the Test Users in the DEL Collection by Using Wikipedia as External Corpus, Statistically Significant Differences between our Methods and *RelM*, *ExtRelM*, *CoWM*, *CoTagM*, and *TagTM* are Indicated by *r*, *e*, *w*, *c*, and *t*, Respectively

Group U50			
	MAP	**NDCG**	**MRR**
LanM	0.0163	0.0309	0.0184
RelM	0.0205	0.0376	0.0222
ExtRelM	0.0211	0.0501	0.0232
CoWM	0.0674 $^{r, e}$	0.0975 $^{r, e}$	0.0779 $^{r, e}$
CoTagM	0.0557 $^{r, e}$	0.086 $^{r, e}$	0.0581 $^{r, e}$
TagTM	0.1525 $^{r, e, w, c}$	0.1924 $^{r, e, w, c}$	0.2009 $^{r, e, w, c}$
EnUWEM	0.2149 $^{r, e, w, c, t}$	0.2586 $^{r, e, w, c, t}$	0.2663 $^{r, e, w, c, t}$
EnUTM	0.2688 $^{r, e, w, c, t}$	0.3111 $^{r, e, w, c, t}$	0.3067 $^{r, e, w, c, t}$
Group U500			
	MAP	**NDCG**	**MRR**
LanM	0.0167	0.0283	0.0203
RelM	0.0221	0.0464	0.0256
ExtRelM	0.0242	0.0468	0.0263
CoWM	0.0886 $^{r, e}$	0.1195 $^{r, e}$	0.0993 $^{r, e}$
CoTagM	0.0249	0.0549 $^{r, e}$	0.0294
TagTM	0.1655 $^{r, e, w, c}$	0.2036 $^{r, e, w, c}$	0.203 $^{r, e, w, c}$
EnUWEM	0.2524 $^{r, e, w, c, t}$	0.285 $^{r, e, w, c, t}$	0.2826 $^{r, e, w, c, t}$
EnUTM	0.2546 $^{r, e, w, c, t}$	0.2955 $^{r, e, w, c, t}$	0.2864 $^{r, e, w, c, t}$
Group UG500			
	MAP	**NDCG**	**MRR**
LanM	0.019	0.0349	0.0193
RelM	0.0305	0.0662	0.0316
ExtRelM	0.0319	0.0674	0.0333
CoWM	0.0916 $^{r, e}$	0.1246 $^{r, e}$	0.1015 $^{r, e}$
CoTagM	0.0485 $^{r, e}$	0.0782 $^{r, e}$	0.556 $^{r, e}$
TagTM	0.2004 $^{r, e, w, c}$	0.2405 $^{r, e, w, c}$	0.2528 $^{r, e, w, c}$
EnUWEM	0.243 $^{r, e, w, c, t}$	0.2922 $^{r, e, w, c, t}$	0.2793 $^{r, e, w, c, t}$
EnUTM	0.254 $^{r, e, w, c, t}$	0.304 $^{r, e, w, c, t}$	0.3029 $^{r, e, w, c, t}$

As seen from Table 4, three conclusions emerge. First, *EnUWEM* and *EnUTM* which rely on the integration of LDA and WEs both score higher than all personalization methods previously proposed, in all metrics measured in all three groups of test users. Moreover, the difference between our proposed methods and the baseline runs is always significant. We believe that the strong performance of our methods is due to the integration of WEs and topic models, and the fact that our methods use an external knowledge base to enhance user profile generation. This provides evidence that our proposed methods may be more beneficial to personalized QE than the previously widely used LDA-based representations and co-occurrence-based techniques. Secondly, both the *EnUWEM* and *EnUTM* methods demonstrate better performance than the strongest personalized baseline. This shows the flexibility of our query expansion methods. The researchers can choose either method according to their personal needs for QE. We note that *EnUTM* works better than *EnUWEM*. As pointed out at the start of Section 5, *EnUWEM* uses only the EUPC model to weight the word representations produced by WEs. Instead *EnUTM* fully exploits the advantages of the integration of the two semantic models.

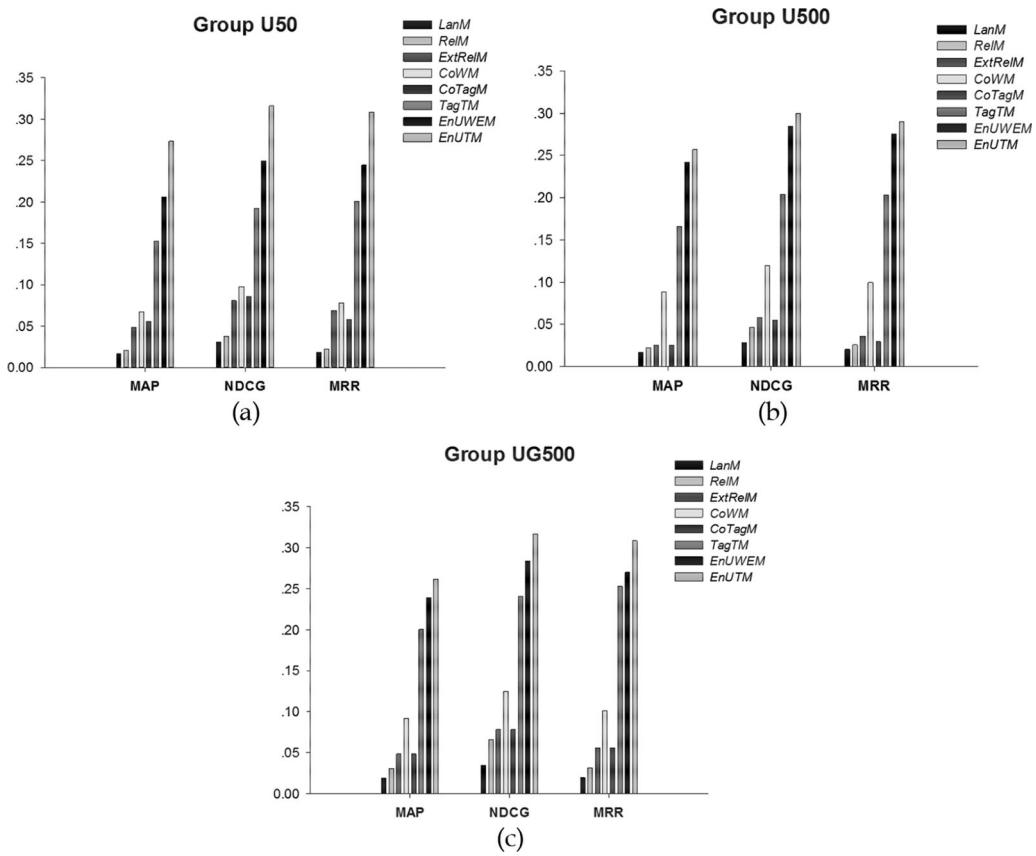

Fig. 5. Results on the test users in the DEL collection by using CLEF News as external corpus.

This also shows that there remains room for improvement in the WEs-based QE method. By treating the documents as a pseudo-aligned corpus, we obtain very good performance. The highest improvement over the best performing run reaches 76.3 percent (in terms of the *EnUTM* method with the MAP metric when compared to *TagTM*). Finally, the improvements brought by our methods are quite consistent across the three different groups of test users. This result confirms that our method works well for users with small and moderate amounts of historical usage information and for those with a rich set of historical data. The improvements obtained in less active users are greater than those obtained by more active users. This shows that our proposed methods are particularly beneficial to those "cold-start" users, who have limited previous interactions with the system.

6.2.2 Performance in the DEL Collection, Using CLEF News as External Corpus

We then compare the performance of different methods using the *DEL* collection by using CLEF News as the external corpus. The results are presented in Fig. 5. It can be seen from the figure that the performance of our proposed methods *EnUWEM* and *EnUTM* is quite stable, as using a different external corpus still achieves statistically significant improvements over all non-personalized and personalized baselines. When compared to using Wikipedia as the external corpus, the improvements using CLEF News are larger. The possible reason is that an external corpus is likely to be a better source of expansion terms if it has very similar topic coverage as the target corpus. Further investigation with respect to concept density will be needed. It is also shown that the *ExtRelM* method works consistently better than the *RelM* and *LanM* baselines. This demonstrates that techniques which explore external corpora to obtain relevant documents work better than methods which simply use the top-ranked documents. These results are consistent with previous research [14].

6.2.3 Performance in the BIB Collection

In order to verify our results on a different test collection, we repeat the experiments using the *BIB* collection. Fig. 6 shows the comparison of our methods *EnUWEM* and *EnUTM* to the methods with different external corpora. Similar to the results of the *DEL* collection, our methods outperform all the non-personalized and personalized baseline methods for all metrics. This time using CLEF News as external corpus shows an even greater improvement than those in the *DEL* collection. Based on all results obtained, we can conclude that our methods outperform all the compared methods on all adopted metrics for both the *DEL* and *BIB* collections using different external corpora. The experimental results demonstrate the advantage and effectiveness of our methods on two different folksonomy datasets.

6.2.4 Effectiveness of the Topical Weighting Scheme

In Section 5.1, when we describe our WEQE method, we weight the word vectors with the posterior estimation of word-topic destributions calculated from the EUPC model. The idea is that the weights will assign more importance to words bearing more information during the generative process. Clearly other alternatives exist, such as *tf*-based weighting, *idf*-based weighting etc [29]. The main

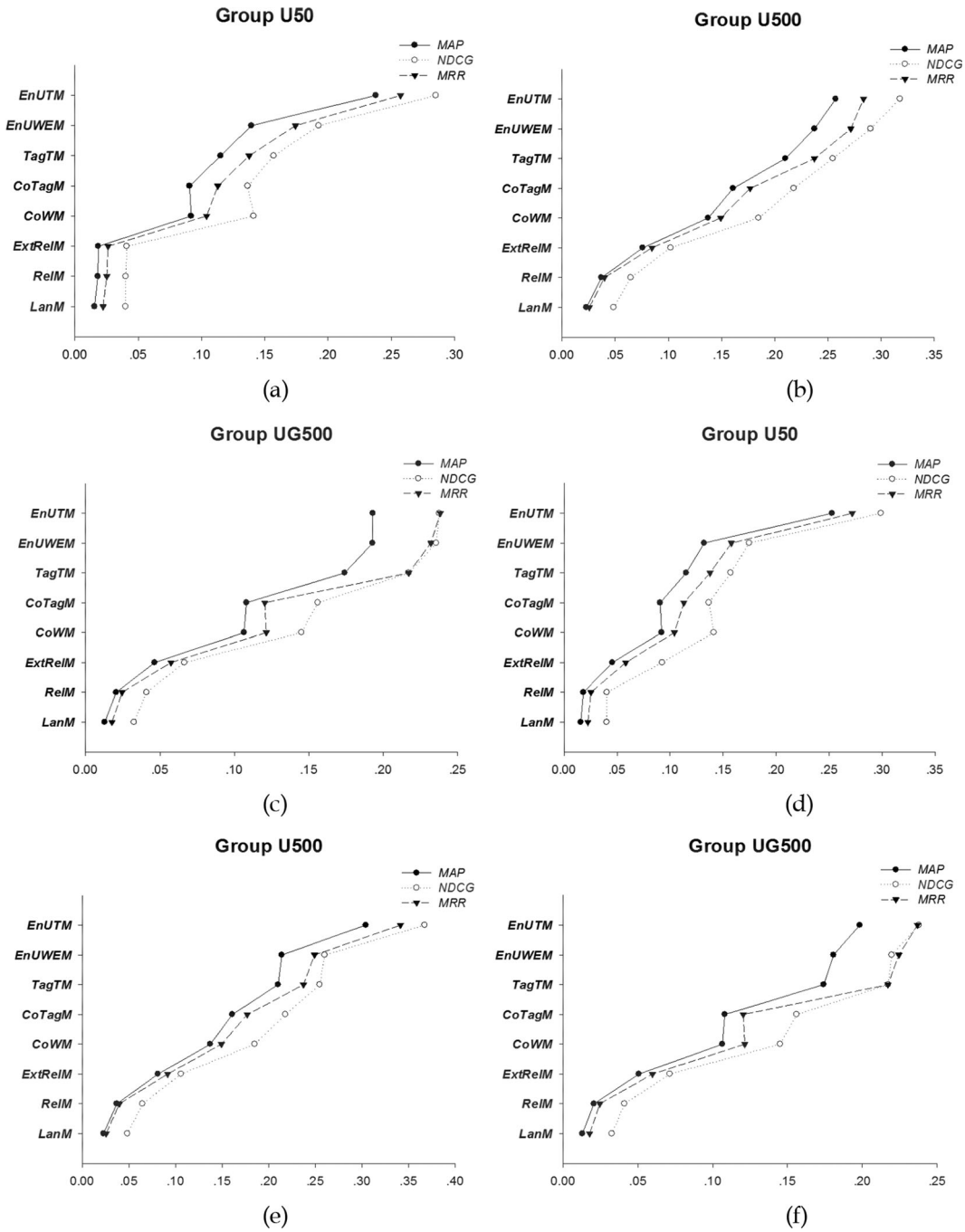

Fig. 6. Results on the test users in the BIB collection, (a)-(c) are results obtained by using Wikipedia as external corpus and (d)-(f) are results obtained by using CLEF News as external corpus.

consideration for using the topical weight is that we want to reflect the multi-aspect of a word and hopefully capture the multiple correlations among words as well as their contexts. To test this assumption, we compare our topical weighting scheme with the *tf*-based (denoted as *EnUWEM-TF*) and *idf*-based (denoted as *EnUWEM-IDF*) weighting schemes. The results are shown in Fig. 7. For consciseness, we only report MAP on the DEL collection using Wikipedia as external corpus, as our experiments show that other test collections and other evaluation metrics all produced similar results. From the figure we can clearly see that *EnUWEM* model works consistently better than the other alternatives. This confirms the effectiveness of our proposed weighting scheme when compared to other simpler weighting schemes. The result also means that using topical weighting can result in a better word representation, and reflect its "truer" semantics.

6.2.5 Effectiveness of the EUPC Model

To verify the effectiveness of our user profile generation model, in particular the effectiveness of integrating word embeddings in topic models, we include a simpler comparison method that uses LDA only for personalized QE. This method uses bilingual LDA [30] to build user profiles and then uses the TQE method for personalized search. This method is denoted as *LdaTM*. As in Section 6.2.4, we only report the MAP results on the DEL collection, using Wikipedia as external corpus. The results are shown in Fig. 8. As can been seen from the figure, our proposed methods *EnUWEM*

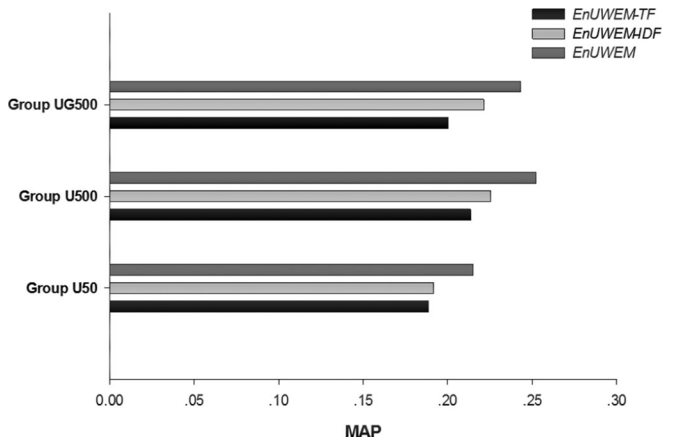

Fig. 7. Effectiveness of the topical weighting scheme.

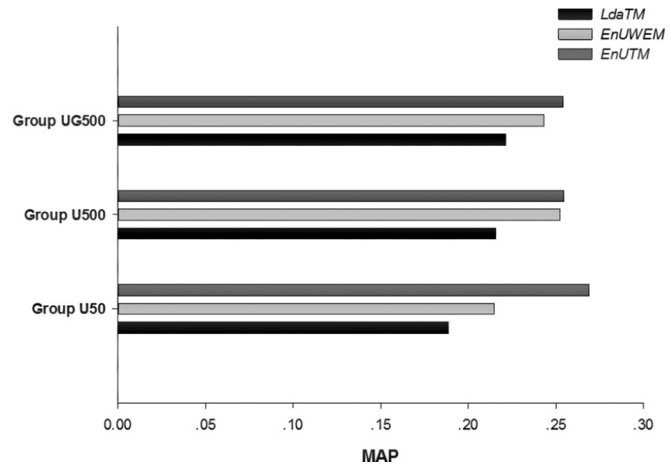

Fig. 8. Effectiveness of the EUPC model.

and *EnUTM* work consistently better than the *LdaTM* method for all three groups of users. This confirms that integrating word embeddings and topic models together can produce better relationships between terms and/or words, as hypothesised at the beginning of the paper.

7 Conclusion

In this paper we study personalized search through enhanced user profiles and personalized query expansion utilizing folksonomy data. We propose a novel model to build enriched user profiles. Our model integrates the current state-of-the-art text representation learning framework, known as word embeddings, with topic models in two groups of pseudo-aligned documents between user annotations and documents from the external corpus. Based on these enhanced user profiles, we then present two novel QE techniques. The first technique approaches the problem by using topical weights-enhanced word embeddings to select the best possible expansion terms. The second technique calculates the topical relevance between the query and the terms inside a user profile. The proposed models performed well on two real-world social tagging datasets produced by folksonomy applications, delivering statistically significant improvements over non-personalized and personalized representative baseline systems. We also show that our method works well for users with small, moderate and rich amounts of historical usage information. In future research, we aim to investigate incorporating more information into the latent semantic model in order to capture more accurate user profiles. Future work will also include the evaluation of different similarity models and weighting schemes to be used in our models.

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, Scientific Research Fund of Hunan Provincial Education Department of China under Grant No. 16K030, Hunan Provincial Natural Science Foundation of China under Grant No. 2017JJ2101, Scientific Research Foundation for the Re-turned Overseas Chinese Scholars, State Education Ministry, China under grant No. [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 13/RC/2106) and is co-funded under the European Regional Development Fund. The authors would like to thank the anonymous reviewers for their contribution to the paper. Dong Zhou is the corresponding author.

References

[1] M. R. Bouadjenek, H. Hacid, and 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.*, vol. 56, pp. 1–18, 2016.

[2] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu, "Exploring folksonomy for personalized search," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Development Inf. Retrieval*, 2008, pp. 155–162.

[3] G. Smith, Folksonomy: social classification, Blog article, Aug. 2004, http://atomiq.org/archives/2004/08/folksonomy_social_classification.html

[4] M. Bertier, R. Guerraoui, V. Leroy, and A.-M. Kermarrec, "Toward personalized query expansion," in *Proc. 2nd ACM EuroSys Workshop Soc. Netw. Syst.*, 2009, pp. 7–12.

[5] C. Biancalana, F. Gasparetti, A. Micarelli, and G. Sansonetti, "Social semantic query expansion," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 4, pp. 1–43, 2013.

[6] D. Zhou, S. Lawless, and V. Wade, "Improving search via personalized query expansion using social media," *Inf. Retrieval*, vol. 15, no. 3/4, pp. 218–242, 2012.

[7] C. Biancalana and A. Micarelli, "Social tagging in query expansion: A new way for personalized web search," in *Proc. Int. Conf. Comput. Sci. Eng.*, 2009, pp. 1060–1065.

[8] M. Bender, et al., "Exploiting social relations for query expansion and result ranking," in *Proc. IEEE 24th Int. Conf. Data Eng. Workshop*, 2008, pp. 501–506.

[9] E. Agichtein, E. Brill, S. Dumais, and R. Ragno, "Learning user interaction models for predicting web search result preferences," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Development Inf. Retrieval*, 2006, pp. 3–10.

[10] R. Das, M. Zaheer, and C. Dyer, "Gaussian LDA for topic models with word embeddings," in *Proc. 53rd Annu. Meet. Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Language Process. Asian Federation Natural Language Process.*, 2015, pp. 795–804.

[11] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun, "Topical word embeddings," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2418–2424.

[12] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Comput. Surv.*, vol. 44, no. 1, pp. 1–50, 2012.

[13] V. Lavrenko and W. B. Croft, "Relevance based language models," in *Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Development Inf. Retrieval*, 2001, pp. 120–127.

[14] F. Diaz and D. Metzler, "Improving the estimation of relevance models using large external corpora," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Development Inf. Retrieval*, 2006, pp. 154–161.

[15] L. Abouenour, K. Bouzouba, and P. Rosso, "An evaluated semantic query expansion and structure-based approach for enhancing Arabic question/answering," *Int. J. Inf. Commun. Technol.*, vol. 3, no. 3, pp. 37–51, 2010.
[16] P.-A. Chirita, C. S. Firan, and W. Nejdl, Personalized query expansion for the web, in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Development Inf. Retrieval*, 2007, pp. 7–14.
[17] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma, "Query expansion by mining user logs," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 4, pp. 829–839, Jul.-Aug. 2003.
[18] I. Ruthven, "Re-examining the potential effectiveness of interactive query expansion," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Development Inf. Retrieval*, 2003, 213–220.
[19] M. R. Bouadjenek, H. Hacid, M. Bouzeghoub, and J. Daigremont, "Personalized social query expansion using social bookmarking systems," in *Proc. 34th Int. ACM SIGIR Conf. Res. Development Inf. Retrieval*, 2011, pp. 1113–1114.
[20] D. Vallet, I. Cantador, and J. M. Jose, "Personalizing web search with folksonomy-based user and document profiles," in *Advances in Information Retrieval*. Berlin, Germany: Springer, 2010, pp. 420–431.
[21] M. Noll and C. Meinel, "Web search personalization via social bookmarking and tagging," in *The Semantic Web*, K. Aberer, et al., Eds. Berlin, Germany: Springer, 2007, pp. 367–380.
[22] Q. Wang and H. Jin, "Exploring online social activities for adaptive search personalization," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, 2010, pp. 999–1008.
[23] M. R. Bouadjenek, H. Hacid, and M. Bouzeghoub, "Sopra: A new social personalized ranking function for improving web search, in *Proc. 36th Int. ACM SIGIR Conf. Res. Development Inf. Retrieval*, 2013, pp. 861–864.
[24] M. R. Bouadjenek, H. Hacid, M. Bouzeghoub, and A. Vakali, "Using social annotations to enhance document representation for personalized search," in *Proc. 36th Int. ACM SIGIR Conf. Res. Development Inf. Retrieval*, 2013, pp. 1049–1052.
[25] Y. Cai and Q. Li, "Personalized search by tag-based user profile and resource profile in collaborative tagging systems," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, 2010, pp. 969–978.
[26] H. Xie, et al., "Personalized search for social media via dominating verbal context," *Neurocomputing*, vol. 172, pp. 27–37, 2016.
[27] H. Xie, et al., "Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy," *Inf. Process. Manage.*, vol. 52, no. 1, pp. 61–72, 2016.
[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
[29] I. Vulić and M.-F. Moens, "Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings," in *Proc. 38th Int. ACM SIGIR Conf. Res. Development Inf. Retrieval*, 2015, pp. 363–372.
[30] I. Vulić, W. De Smet, and M.-F. Moens, "Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora," *Inf. Retrieval*, vol. 16, no. 3, pp. 331–368, 2013.
[31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Advances Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[32] D. Ganguly, J. Leveling, and G. F. Jones, Topical Relevance Model, in *Information Retrieval Technology*. Y. Hou, et al., Eds. Berlin, Germany: Springer, 2012, pp. 326–335.
[33] J. Mitchell and M. Lapata, "Vector-based Models of Semantic Composition," in *Proc. ACL: HLT*, 2008, pp. 236–244.
[34] A. Zubiaga, A. P. Garcia-Plaza, V. Fresno, and R. Martinez, "Content-based clustering for tag cloud visualization," in *Proc. Int. Conf. Advances Soc. Netw. Anal. Mining*, 2009, pp. 316–319.
[35] A. Zubiaga, V. Fresno, R. Martinez, and A. P. Garcia-Plaza, "Harnessing folksonomies to produce a social classification of resources," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 8, pp. 1801–1813, Aug. 2013.
[36] C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Proc. 10th Int. Conf. Inf. Knowl. Manage.*, 2001, pp. 403–410.

Dong Zhou received the PhD degree from the University of Nottingham, in 2009, Unided Kingdom. He is an associate professor at the Key Laboratory of Knowledge Processing and Networked Manufacturing & School of Computer Science and Engineering, Hunan University of Science and Technology, China. He worked as a research fellow in Trinity College Dublin, Ireland, from 2008 to 2012 in the Centre for Next Generation Localization. His current research interests include information retrieval, personalization, natural language processing, and data mining.

Xuan Wu received the BSc degree in network engineering from the Hunan University of Science and Technology, China, in 2015. She is currently a member of the Key Laboratory of Knowledge Processing and Networked Manufacturing at the Hunan University of Science and Technology. Her research interests include social networks and information retrieval.

Wenyu Zhao received the BSc degree in network engineering from the Hunan University of Science and Technology, China, in 2015. She is currently a member of the Key Laboratory of Knowledge Processing and Networked Manufacturing at the Hunan University of Science and Technology. Her research interests include information retrieval and personalized search.

Séamus Lawless received the BSc and PhD degrees from the same institute, in 2003 and 2009, respectively. He is an assistant professor at the School of Computer Science and Statistics, Trinity College Dublin, Ireland. His research interests are in the areas of information retrieval, information management, and digital humanities with a particular focus on adaptivity and personalization.

Jianxun Liu received the PhD degree in computer application technology from Shanghai Jiaotong University, China, in 2003. He is a professor in the School of Computer Science and Engineering, Hunan University of Science and Technology, China. He is the associate director of the Key Laboratory of Knowledge Processing and Networked Manufacturing. He has been selected for the Program for New Century Excellent Talents in University (NCET), China. His research interests include service computing, workflow management, and application.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.