EAI Endorsed Transactions
on Collaborative Computing

Research Article EAI.EU

An Augmented User Model for Personalized Search in Collaborative Social Tagging Systems

Wenyu Zhao[1], Dong Zhou[1,*] , Xuan Wu[1], S éamus Lawless[2] and Jianxun Liu[1]

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

Abstract

Alongside the enormous volume of user-generated content posted to World Wide Web, there exists a thriving demand for search personalization services. To provide personalized services, a user model is usually required. We address the setting adopted by the majority of previous work, where a user model consists solely of the user's past information. We construct an augmented user model from a number of tags and documents. These resources are further processed according to the user's past information by exploring external knowledge base. A novel generative model is proposed for user model generation. This model utilizes recent advances in neural language models such as Word Embeddings with latent semantic models such as Latent Dirichlet Allocation. We further present a new query expansion method to facilitate the desired personalized retrieval. Experiments conducted on two real-world collaborative social tagging datasets show that our proposed methods outperform state-of-the-art methods.

Keywords: Personalized Social Search, Collaborative Social Tagging Systems, Latent Dirichlet Allocation, Neural Language Model, Query Expansion

Received on 30 September 2017, accepted on 06 October 2017, published on 09 October 2017

Copyright © 2017 Wenyu Zhao *et al.*, licensed to EAI. This is an open access article distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/3.0/), which permits unlimited use, distribution and reproduction in any medium so long as the original work is properly cited.

doi: 10.4108/eai.9-10-2017.154549

*Corresponding author. Email:dongzhou1979@hotmail.com

1. Introduction

The amount of digital content online has increased exponentially recently. The use of personalized Web search systems has become crucial in retrieving relevant information. Such system fetches relevant information that are most correlated to an individual user rather than only to the issued query [1, 2]. Recording of the individual's interests and past behaviors in user models has been widely adopted. Subsequently the information inside the user model can be used for query and/or results personalization.

With this increasing volume of digital content comes an increasing number of collaborative social tagging Websites for Web pages and documents. Collaborative social tagging systems like *del.icio.us*[1] and *BibSonomy*[2], etc., have become more and more popular. The tags and documents added by different users to the platforms are closely linked to that individual and their interests, providing abundant information for constructing more rigid and characteristic user models. Therefore, constructing user models from social tagging systems has the potential to be instrumental for personalized search. In the social tagging platform, users are freely to choose whatever words/terms to be used in tagging. This behavior makes the information search process

[1] http://www.delicious.com/
[2] http://www.bibsonomy.org

Table 1. Basic notations used in the paper

Symbol	Meaning	Symbol	Meaning
$\mathcal{U}$	finite sets of users	σ	deviation of the normal distribution
$\mathcal{D}$	finite sets of documents	K	numbers of topics
$\mathcal{T}$	finite sets of tags	α	the parameters of topic Dirichlet prior
$\mathcal{A}$	a ternary relation, elements are tags	β	the parameters of word Dirichlet prior
u	a user	θ	multinomial distribution of topics
t	a tag	φ	multinomial distribution of words
d	a document	E	dimensions of word embeddings
w	a word/term	$w_{j,i}$	ith word in document
$\mathcal{A}^u$	the set of annotations of a user	$f_{j,i}^e$	dimension e of the embedding of word $w_{j,i}$
$\mathcal{D}^u$	a user's set of documents	$z_{j,i}$	topic associated with ith word in document d_j
$\mathcal{T}^u$	the tag vocabulary of a user	N_d	number of words in a document
$\mathcal{D}_{exter}$	an external corpus	μ_z	mean of Log-normal distribution of retrieval scores for topic z
$term^{\mathcal{D}^u}$	terms extracted from documents annotated by a user	σ_z	deviation of Log-normal distribution of retrieval scores for topic z
$\mathcal{D}_{exter}^u$	a user's set of external documents	$n_{j,k}$	the number of times that topic k sampled from document d_j
$term^{\mathcal{D}_{exter}^u}$	set of terms extracted from a user's set of external documents	$v_{k,w_{j,i}}$	the number of times $w_{j,i}$ generated by topic k
q	a original query	d_b	relevant document
$q^{\mathcal{T}^u}$	a query containing the concatenated tags of a user	n	The number of independent query words
$Q^{\mathcal{D}^u}$	a query extracted from a document d^u that a user tagged	H	a hidden model
λ	number of top terms extracted from $\mathcal{D}^u$	M	total number of documents
γ	number of documents extracted from $\mathcal{D}_{exter}^u$	$topic_k$	a particular topic learnt
μ	the mean of the normal distribution	δ	number of top terms selected from user models

even more difficult than normal web search systems. To deal with this problem, personalized search results re-ranking [3-6] and personalized query expansion [7-9] have been widely adopted.

However, there are several drawbacks in the process of searching in such type of systems, including the following. i) In the approaches that a user model contains potential expansion terms, past researchers used relationships between tags and lexical matching methods between terms and queries. For the most of the cases, tags can not be viewed as accurate summarization of documents, henceforth the search experiences are somewhat depressed [10]. Moreover, lexical matching may miss some latent semantic information exhibited in the user model. ii) All the previously proposed personalized search methods require a user's past click/browse information stored in a user model. However, we argue that using this information alone is not sufficient. In some cases, a user may have clicked and/or browsed only a few documents. It is relatively hard to personalized search with this little usage information to hand.

To handle these limitations, we construct an augmented user model from a number of tags and documents. These resources are further processed according to the user's past information by exploring external knowledge base. A novel generative model is proposed for user model generation. This model leverages recent advances in neural language models such as Word Embeddings (WE) [11] with the traditional Latent Dirichlet Allocation (LDA) model [12]. We learn latent topics that generate word embeddings and words simultaneously. Based on the topics learnt, we further propose a novel topical query expansion model to be used in social tagging systems. In this model, queries are expanded not only by their lexical similarity with the potential terms, but are also based on their topical relevancy. Observing the obtained evaluation results from two social datasets sourced from two real-world platforms, we can find that the personalized search methods provide very good performance improvements over various baseline methods.

Our contribution in the current paper are: i) We introduce augmented user profiles by exploiting an external knowledge base to perform personalized search in a novel

way. ii) We suggest and evaluate a novel generative model that leverages recent advances in neural language models with latent semantic models.

2. Related Work

There exist sufficient researches in personalized search [1, 2]. These can be roughly allocated into two categories. The first one is known as results processing. This is usually done with results re-ranking by re-ordering retrieved results using the information from a user model. [13]. Another is query expansion [14]. In this category, new terms selecting from a user model can be utilized to expand the original query, or terms inside the initial query can be re-weighted according to the user model. [15].

The problem studied in the current paper falls in one of the above two strategies. Tags and documents crawled from a social tagging system can be used to construct a test collection. This collection can be further utilized to advance the research in search personalization. For example, tags and documents can be employed to automatically learn an individual's preferences. The retrieval results can be personalized based on the topical relevance between documents and information from the user model [3]. Signals from multiple rather than single social systems can be used for search personalization [16]. Bouadjenek et al. [4] presented an enhanced document representation based on user relationships to re-rank documents on a social platform. Cai et al. [17] treated the relevance as fuzzy satisfaction between users and queries. However, if the relevant documents cannot be returned in the retrieval list, the strategy has no way to fetch more relevant results.

Another strategy expands a user's issued query with potential terms from user models. Relationships between tags have been used for personalized query expansion. If a tag appears in a query, the most related tags will be selected from the user model to expand the query [10]. Researchers also considered using lexical matching methods such as co-occurrence-based method to expand the query based on a user's past information [15]. Recently, a query expansion method for personalization has been proposed [8], which is state-of-the-art. This method captures term relationships through a Tag-Topic model. Mutual information between the terms is then utilized to choose the potential expansion terms.

All of the above systems consider building user models from his/her past information only. In contrast, in this paper we explore an external knowledge base to build augmented user models. We also propose a novel topical query expansion model for personalization.

3. User Model Generation

We now define the research problem studied in the current paper. Subsequently we describe the user model generation process. Formally, data in social tagging systems can be represented by $\mathcal{P} \coloneqq (\mathcal{U}, \mathcal{D}, \mathcal{T}, \mathcal{A})$. The elements in the ternary relation $\mathcal{A} \subseteq \mathcal{U} \times \mathcal{D} \times \mathcal{T}$ are called annotations, or bookmarking activities performed by different users. $\mathcal{A}^u \coloneqq \{(t,d)|u \in \mathcal{U}, d \in \mathcal{D}, t \in \mathcal{T}\}$ represents a user's annotations. $\mathcal{D}^u \coloneqq \{d|(t,d) \in \mathcal{A}^u\}$ represents all documents annotated by a user u. $\mathcal{T}^u \coloneqq \{t|(t,d) \in \mathcal{A}^u\}$ represents the total tags used by a user. $term^{\mathcal{D}^u} \coloneqq \{w|w \in \mathcal{D}^u\}$ represents all terms from $\mathcal{D}^u$, w is a word/term extracted from $\mathcal{D}^u$. Similarly, $term^{\mathcal{D}^u_{exter}} \coloneqq \{w|w \in \mathcal{D}^u_{exter}\}$ represents all terms extracted from $\mathcal{D}^u_{exter}$. $\mathcal{D}^u_{exter}$ represents all external documents extracted from an corpus $\mathcal{D}_{exter}$ different from $\mathcal{D}$. In table 1, we list all the definition of basic notations used in our paper.

Algorithm 1 Gen*erative process for Enhanc*ed User Model Generation

Require: the total tags used by a user $\mathcal{T}^u$

Require: all documents annotated by 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 $\mathcal{T}^u \cup \mathcal{D}^u \cup \mathcal{D}^u_{exter}$ $(\mathcal{D}')$

1. **for** $k \in [1, K]$
2. draw mixture components $\varphi_k \sim Dirichlet(\beta)$
3. **for** d_j in $\mathcal{D}'$
4. draw mixture proportion $\theta_j \sim Dirichlet(\alpha)$
5. **for** $w_i \in [1, N_{d_j}]$
6. draw topic index $z_{j,i} \sim Multnomial(\theta_{d_j})$
7. draw term for word $w_{j,i} \sim Multnomial(\varphi_{z_{j,i}})$
8. for each dimension of the embedding of $w_{j,i}$, draw $f^e_{j,i} \sim \mathcal{N}(\mu^e_{z_{j,i}}, \sigma^e_{z_{j,i}})$

In the problem studied here, we construct a user model contains a number of words/terms $\{w_1, w_2 \ldots w_n\} \in term^{\mathcal{D}^u} \cup term^{\mathcal{D}^u_{exter}} \cup \mathcal{T}^u$. If a user submits a query, potential expansion terms will be selected from a sorted list of terms.

The user model generation has two steps: external documents fetch and user model construction. We augment a user's past information in step 1. All tags t in $\mathcal{T}^u$ are joined to form a query $q^{\mathcal{T}^u}$. We iterate through d in $\mathcal{D}^u$ to extract high weighted terms to form a number of queries $Q^{\mathcal{D}^u}$ (inverted document frequency used here and we extract top λ terms). Then we issue queries in $q^{\mathcal{T}^u} \cup Q^{\mathcal{D}^u}$ to an external corpus $\mathcal{D}_{exter}$ to fetch γ number of documents to form $\mathcal{D}^u_{exter}$.

In step 2, we integrate $\mathcal{T}^u$, $\mathcal{D}^u$ and $\mathcal{D}^u_{exter}$ into a novel generative model. In this model, multinomial distribution of topics of each single document can be easily acquired. The procedure is described in the remaining of this section.

It is well known that the LDA model can mine the thematic structure of documents. Recently, WE has played an increasingly vital role in building continuous word vectors based on their context in a corpus. There are also some attempts to integrate LDA with WE for different purposes [18, 19]. Inspired by those works, a novel generative model for user model generation is presented in

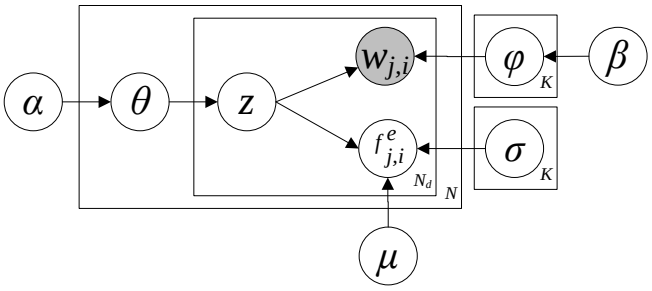

Fig 1. Plate notation of the EUMG model

this paper. We named this model enhanced user model generation (EUMG).

To jointly model words and word embeddings produced by WE, EUMG learns a shared latent topic space to generate words in documents and corresponding word embeddings. The WE are all pre-trained, together with documents as input to the model. Skip-Gram model [11] is utilized here before running our model to learn WE. A normal distribution is used for WE to learn latent topics from the documents as well as the words. With the WE and documents trained by the Skip-Gram model, the generation process of the EUMG model can be summarized as in Algorithm 1.

In Algorithm 1, the mean and deviation of the normal distribution are defined as μ and σ respectively. The parameters of a topic Dirichlet prior and word Dirichlet prior are defined as α and β. θ_j is the multinomial topic distribution of document d_j. $f^e_{j,i}$ are word embeddings. The number of latent topics and dimensions for WE are both fixed. Both the words and WE determine posterior distribution of topics.

In the proposed model, Gibbs Sampling is used to solve the intractable inference problem. A conjugate prior is used here. We also integrate out θ and φ. The conditional distribution $p(z_{j,i} = k)$ has to be calculated in sampling. Specifically, the topic is chosen from the following equation for each word:

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

The amount of times that topic k is added up in $n_{j,k,\neg i}$ (from multinomial distribution of the document j). Note that the present $z_{j,i}$ is not included. The amount of times $w_{j,i}$ is generated by topic k is added up in $v_{k,w_{j,i},\neg}$. The present $w_{j,i}$ is not included. The summation over all values of the variable is denoted by a dot. E is the dimensions of word embeddings. The posterior estimate of θ and φ can then be easily obtained. The plate notation of EUMG model is shown in Fig.1.

Table 2. Statistics for two test collections

	Users	Documents	Tags
del.icio.us	259,511	131,283	137,870
bibsonomy	6,733	208,260	73,647

4. Topical Query Expansion

Next, The output from step 2 of the last section can be utilized to build a query expansion model that ranks terms from the user model to be added to the query. We only layout the key steps in this section because of space constrains.

We assume there exists a query $q = \{w_a\}_{a=1}^n$, where $\{w_a\}_{a=1}^n$ denotes n independent query words. We approximate the probability of q generating w from a hidden model H is by: (see also [20, 21]):

$$P(w|H) \approx P(w|q) \tag{2}$$

We further define a number of relevant documents $\{d_b\}_{b=1}^M$. These documents have relationships with both the query and the words in a user model. M represents total number of documents. Associate $\{d_b\}$ into equation (2) leads to:

$$P(w|q) = \sum_{b=1}^{M} P(w|d_b)P(d_b|q) \tag{3}$$

The uniform prior of documents is also put outside of the summation. $P(q)$ has been eliminated here as a uniform prior.

The output form step 2 of the last section (i.e. the documents in the user model) can be treated as $\{d_b\}_{b=1}^M$ in the above equation. In equation (3), w has a direct dependency on d_b and w_a also has a direct dependency on d_b, the assumption is too simplistic. Through the EUMG model, we obtain latent topics. These topics can be used to re-calculate the probability of q generating w:

$$\begin{aligned} & P(w|q) \\ & \propto \frac{1}{M} \sum_{b=1}^{M} \left(\sum_{k=1}^{K} P(w|topic_k)P(topic_k|d_b) \right) \\ & \times \left(\prod_{a=1}^{n} \sum_{k=1}^{K} P(w_a|topic_k)P(topic_k|d_b) \right) \end{aligned} \tag{4}$$

$topic_k$ represents a particular topic learnt. After we obtain the probability scores, we sort all the terms and select the top δ terms as the final expansion query terms.

5. Experiments

5.1. Evaluation Setup

To examine the performance of the user model generation and query expansion methods, we perform the experiments in two datasets which are *del.icio.us* and *bibsonomy*. The first collection (DEL) merges two real-world sub-datasets from a social tagging system *del.icio.us*: *socialbm0311* and *deliciousT140*. Please refer to [22, 23] for details about the two datasets. There are 5,153,720 annotation activities,

Table 3. Evaluation results of various methods on DEL colletion. Statistically significant differences between a method with the best performing non-personalized baseline (*LM+RM-wiki*) and the best performing personalized baseline (*Tag-topic+QE*) are indicated by *, # respectively

	User50				User500		
	MAP	NDCG	MRR		MAP	NDCG	MRR
LM	0.0163	0.0309	0.0184	***LM***	0.0167	0.0283	0.0203
LM+RM	0.0205	0.0376	0.0222	***LM+RM***	0.0188	0.0317	0.0224
LM+RM-wiki	0.0211	0.0501	0.0232	***LM+RM-wiki***	0.0242	0.0468	0.0263
CoTag+QE	0.0557	0.086	0.0581	***CoTag+QE***	0.0249	0.0549	0.0294
Cooccur+QE	0.0674*	0.0975*	0.0779*	***Cooccur+QE***	0.0886*	0.1195*	0.0993*
Tag-topic+QE	0.1525*	0.1924*	0.2009*	***Tag-topic+QE***	0.1655*	0.2036*	0.203*
EUMG+TQE	0.2440*,#	0.2868*,#	0.2980*,#	***EUMG+TQE***	0.2154*,#	0.2592*,#	0.2424*,#
	User100				UserG500		
	MAP	NDCG	MRR		MAP	NDCG	MRR
LM	0.0125	0.0314	0.0136	***LM***	0.019	0.0349	0.0193
LM+RM	0.0132	0.0347	0.0137	***LM+RM***	0.0305	0.0662	0.0316
LM+RM-wiki	0.0225	0.0384	0.0238	***LM+RM-wiki***	0.0319	0.0674	0.0333
CoTag+QE	0.0437	0.084	0.0448	***CoTag+QE***	0.0485	0.0782	0.0556
Cooccur+QE	0.0843*	0.1216*	0.0897*	***Cooccur+QE***	0.0916*	0.1246*	0.1015*
Tag-topic+QE	0.1586*	0.1647*	0.1721*	***Tag-topic+QE***	0.2004*	0.2405*	0.2528*
EUMG+TQE	0.2117*,#	0.2476*,#	0.2377*,#	***EUMG+TQE***	0.2385*,#	0.2897*,#	0.2802*,#

259,511 users, 137,870 tags and 131,283 documents in our merged dataset. The second collection (BIB) comes from *bibsonomy* dumps on 01/01/2017, in which there contains 1,665,190 bookmark activities. Before the experiments, we perform four data preprocessing tasks: (1) we utilize a public parser to parse all web pages. (2) We download all documents from available web pages and remove unprofitable web pages. (3) We use English analyzer, Porter's stemmer and a stopword list for documents and tags. (4) All documents from the corpus are indexed by using Terrier[3] toolkit. Our final dataset consists of 6733 users, 208260 documens (web pages) and 73,647 tags. Table 2 shows the statistics of both test collections.

An external knowledge base is constructed from Wikipedia [4]. This knowledge base contains 4,634,369 documents. It was crawled on 14/08/2014. To evaluate the effects of augmented user models, four sets of users with different size are chosen as test users on del.icio.us collection. This includes: users with no more than 50 annotation activities (User50), users with 50-100 annotation activities (User100), users with 100-500 annotation activities (User500) and users with more than 500 annotation activities (UserG500). Compared with the first collection, the number of users is much smaller than that in first collection, so we only choose three sets of users with different size as test users on bibsonomy collection. It contains: users with no more than 50 annotation activities (User50), users with 50-500 annotation activities (User500) and users with more than 500 annotation activities (UserG500). These sets represent users with small, moderate and rich amounts of past usage information respectively. We randomly choose 200 users from each set. We select 25% of each user's annotations for evaluation, and the remaining 75% are utilized to build the user model.

We follow the evaluation procedure of previous research [3, 8, 16]. If a user u issues a query t, this query is viewed as a personalized query. In this case, relevant documents are the documents annotated by u with t.

We use the evaluation metrics that are typically utilized in Web search evaluation: mean average precision (MAP), which is usually employed to report search accuracy; normalized discounted cumulative gain (NDCG), a natural choice for search engine evaluation; as well as another commonly adopted evaluation metric mean reciprocal rank (MRR). Paired t-test is used for significance evaluation. We set the confidence level at 95%. The average performance is computed for all users in the same set.

The methods proposed in this paper are compared with several query expansion methods. This includes non-personalized methods and personalized baselines. We now describe them in detail.

[3] http://www.terrier.org
[4] http://www.wikipedia.org

LM A language model based retrieval method. This model is quite popular and produces good results before. We use the model described in [24], we use it as the initial results without expanding original query.

LM+RM A relevance model is allowed to expand query based on relevant feedback documents, which involves in the pseudo-relevance feedback in the language model as in [19]. This model was regarded as a non-personalized query expansion baseline.

LM+RM-wiki This is a relevance model which uses Wikipedia to obtain the relevance documents, as in [25]. It is a strong non-personalized baseline for comparison because that we also used external corpus in our approach.

CoTag+QE Many researchers tend to consider the tag-tag relationships. The method utilizes the users performed co-tag activities to expand query [7, 10]. So user models are constructed from tags and their co-tag statistics, which computed by using Jaccard coefficient as in [7].

Cooccur+QE This is a personalized baseline method, utilized by many previous researchers [6, 11]. The method calculates co-occurrence scores between terms from a user model and terms from a query [15].

Tag-topic+QE This method captures term relationships through a Tag-Topic model. Mutual information between the terms is then utilized to choose the potential expansion terms [8]. The highest performing method from [8] is selected here as a strong personalized baseline.

EUMG+TQE finally, our approach uses the EUMG model and the query expansion method proposed in section 4 for personalized search in social tagging systems.

The parameters are described as following. Prior parameters α, β and λ of EUMG model are set to $K/50$, 0.01 and 10 respectively, where K denotes the number of latent topics. The number of topics and the number of terms selected from user models are chosen from [5, 50]. The dimension of numbers and the number of documents retrieved from an exterior corpus are chosen from [10, 100] and [1, 10] respectively. When there are $K = 15$ latent topics and $E = 50$ dimensions for WE, $\gamma = 5$ documents retrieved from an exterior corpus, $\delta = 50$ terms selected from user models to expand the original query, we obtain the highest performance.

5.2. Experimental Results

5.2.1 Performance in the DEL collection

Results conducted on the DEL collection are now fully examined in this section. The overall performance is demonstrated in Table 3, including our new approach presented in the paper together with baseline methods on the test users in four sets. Statistically significant differences between a method with the best performing non-personalized baseline (***LM+RM-wiki***) and the best performing personalized baseline (***Tag-topic+QE***) are indicated by *, $^{\sharp}$ respectively. From the results, we learn that ***LM*** model performs the worst in all different sets of users by using all evaluation metrics. ***LM+RM*** method outperformed the ***LM***, because the method based on the pseudo-relevance feedback documents, which can help discover the relevance between words from feedback documents and queries. ***LM+RM-wiki*** method performed steadily better than ***LM*** and ***LM+RM***. This illustrates the effectiveness of utilizing an external corpus. The results are consistent with previous research [25]. These three methods are surpassed by the three personalized methods with statistically significance. This includes the method ***EUMG+TQE*** proposed in this paper. This shows that terms in the user models can improve the effectiveness of search significantly. Non-personalized query expansion methods only select terms from top documents. There are only limited improvements observed.

The personalized baselines ***CoTag+QE***, ***Cooccur+QE*** and ***Tag-topic+QE*** expand the queries by using the user's historical information only while our approach explores an external knowledge base. We now analyze the results for these four methods. As seen from Table 3, several conclusions can be drawn. First, ***EUMG+TQE*** outperforms the three personalization methods ***CoTag+QE***, ***Cooccur+QE*** and ***Tag-topic+QE***, in all four sets of users by using different metrics. The differences are consistently significant. The possible reason is that we use an external knowledge base in addition to the user's past information to build an augmented user model for personalized search. Secondly, ***EUMG+TQE*** achieves consistent improvements over baseline approaches across four sets of users. The improvements in **User50** are more remarkable than in other sets of test users. In reality, we often face the situation where a user has little interactions with the search platform. Under such circumstances, personalized search experience is usually unsatisfactory. However, with enhanced content, our method can obtain reasonably better results for this set of users. This result also confirms that our approach performs well for users with small or moderate volumes of past information and those with a rich set of historical data. Third, using Wikipedia seems a good choice of the external corpus. The possible reason, as pointed out in [25], is that if an external knowledge base has good coverage of topics, it is more likely that good expansion terms can be selected from it.

5.2.2 Performance in the BIB collection

In this section, to further evaluate the performance of our proposed method, we also performed experiments on the BIB collection. The experimental results describe the performance of our proposed method based on external knowledge base and various baselines on the test users with different groups. Fig.2 shows the performance of our proposed in this paper compared with other baselines of different groups by using different evaluation metrics. It's clear from Fig.2 that on different groups, for all the evaluation metrics, our proposed method achieved better results than all of its competitors, whatever it's the best non-personalized method the best personalized method. This improvement can verify the effectiveness of user models by exploiting external knowledge base again, which is

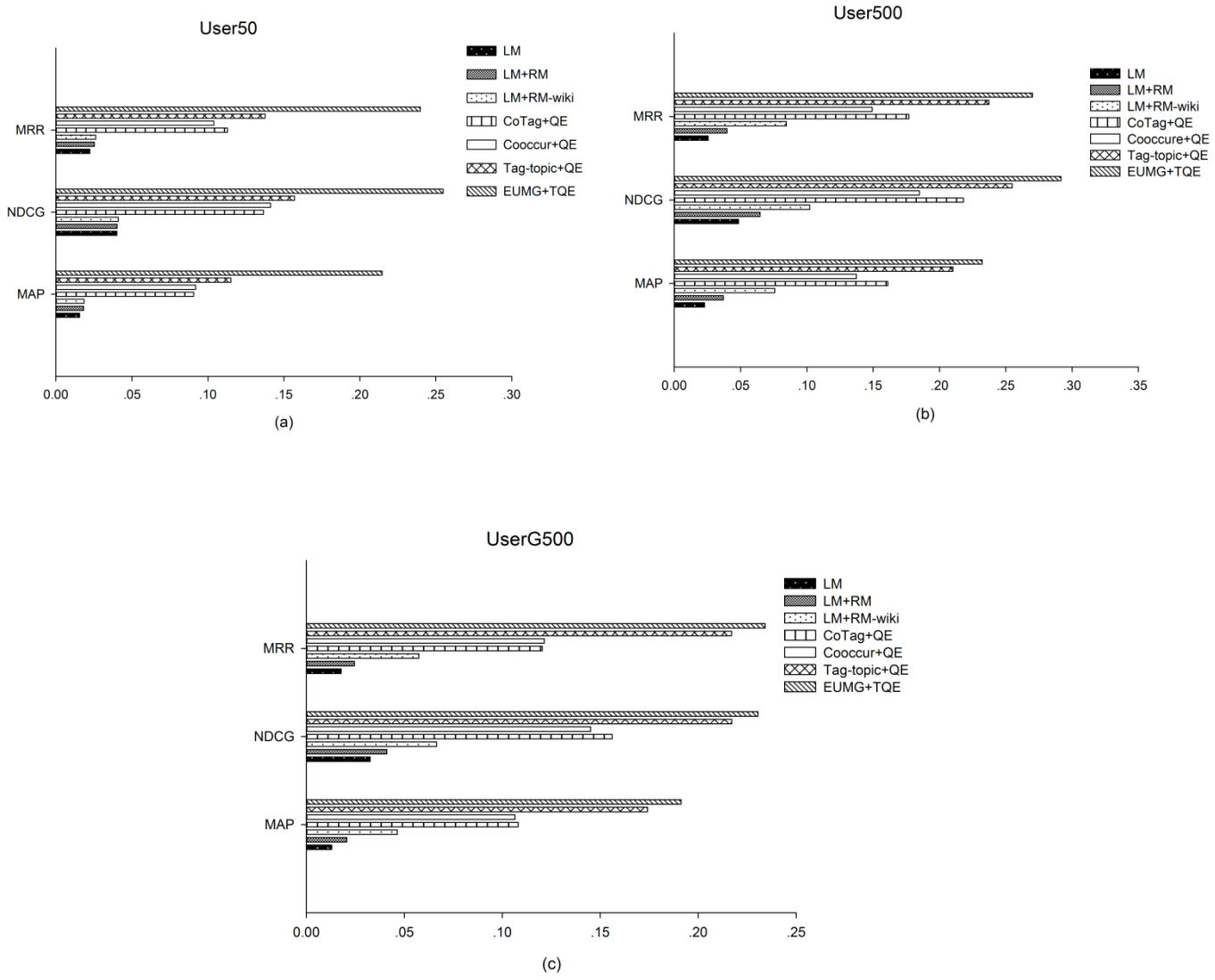

Fig. 2. Results on the test users in the BIB collection

consistent with the results gained by using the DEL collection.

On three groups of test users, the initial method (***LM***) still showed the worst performance. In User500 and UserG500, ***LM+RM-wiki*** showed better performance than ***LM+RM*** with the help of external knowledge base. The improvements demonstrate that external knowledge base can enrich the words information to expand original query.

However, in User50, ***LM+RM-wiki*** and ***LM+RM*** produced similar results. The possible reason is that users had little interacting information with social tagging system and feedback documents had little similar word information to the words in user model. Compared with those approaches that not used external knowledge base, to some extent external knowledge base could enrich the terms expansion sets, but the improvements of non-personalized methods are really limited.

As shown by Fig.2, we can see that three personalized baselines (***CoTag+QE***, ***Cooccur+QE*** and ***Tag-topic+QE***) always outperformed non-personalized methods. Though ***LM+RM-wiki*** method exploited external knowledge base, this methods was based on the feedback documents, which can't supply enough information to term expansion sets. While these personalized baselines utilized the user's historical information to construct user models, such as the relationships between tag-tags, the co-occurrence statistics of tags, the semantic information of tags. This demonstrates that it's quite helpful to make use of user's historical information in user model construction. Apart from this observation, we also can conclude that ***Tag-topic+QE*** method shows the best performance in all personalized baselines. This is due to the factor that the method explored the tag-topic relationships based on tag-topic model, which used LDA model to focus on the semantics of words in user model. It means that the semantic information is very

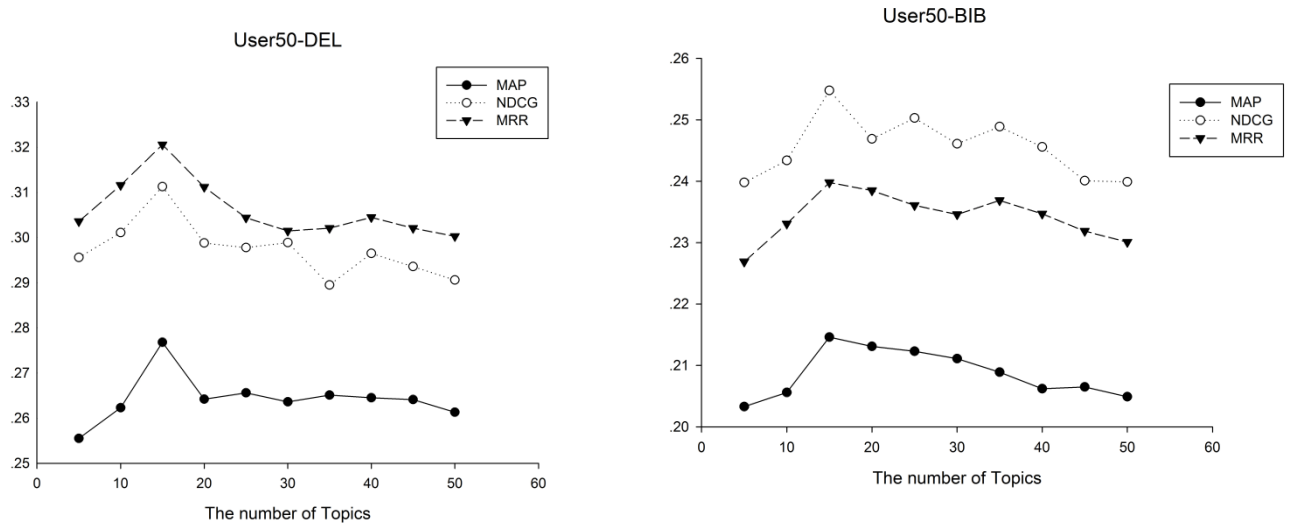

Fig. 3. The performance by varying the number of topics on two corpora

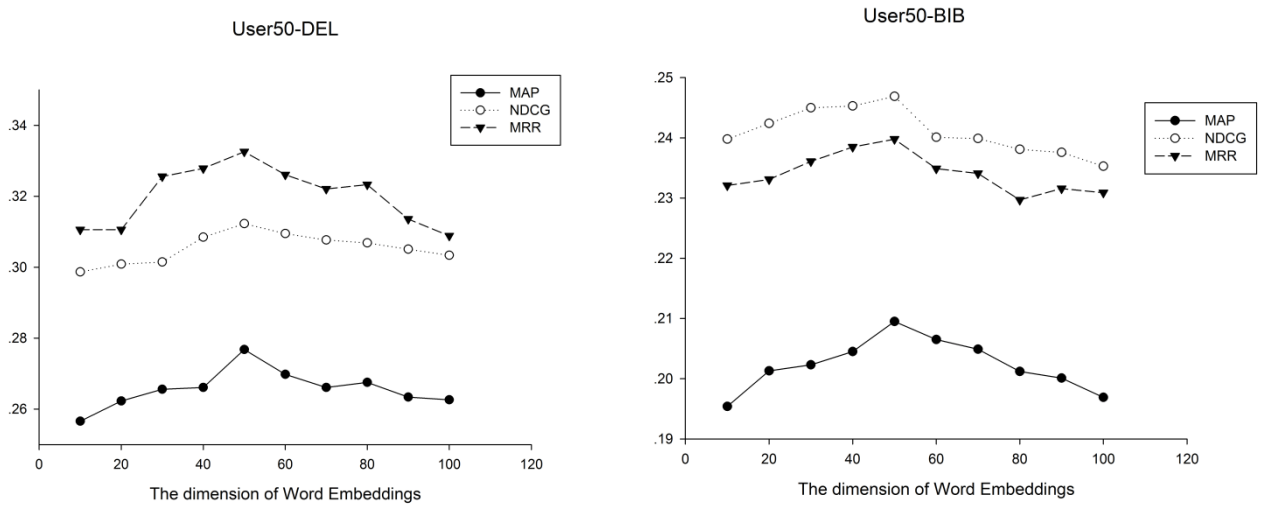

Fig. 4. The performance by varying the number of Word Embeddings on two corpora

beneficial to construct user models for personalized search. A support is that our proposed method have better performance than the best personalized baseline (***Tag-topic+QE***). We consider that there are two reasons. One is that our method combines Word Embeddings with LDA model, which can use the context semantic information of words to enhance the performance of LDA model. The other one is that we exploit external knowledge base to enrich the information in user models.

Based on all results gained by using the DEL collection and BIB collection on different groups of text users, we can conclude that our proposed method in this paper has been proven to be more powerful approach than those methods on the DEL and BIB collection by different evaluation metrics. Those only focusing on using user's historical information to construct user models, or considering the statistical relationships of occurrence rather than real semantic information embedded in words during the process of user model construction. All results show the good effectiveness of our proposed method on two different social tagging systems.

Finally, we examine the optimum number of latent topics, dimensions for WE, the number of documents retrieved from an exterior corpus, the number of terms selected from user models.

5.2.3 Impact of the Number of Topics

To verify the influence of performance about our proposed method by the number of topic, we only test our experiments on User50 from the DEL collection and the BIB collection. As shown in left picture of Fig.3, on the DEL collection, we can observe that the performance of

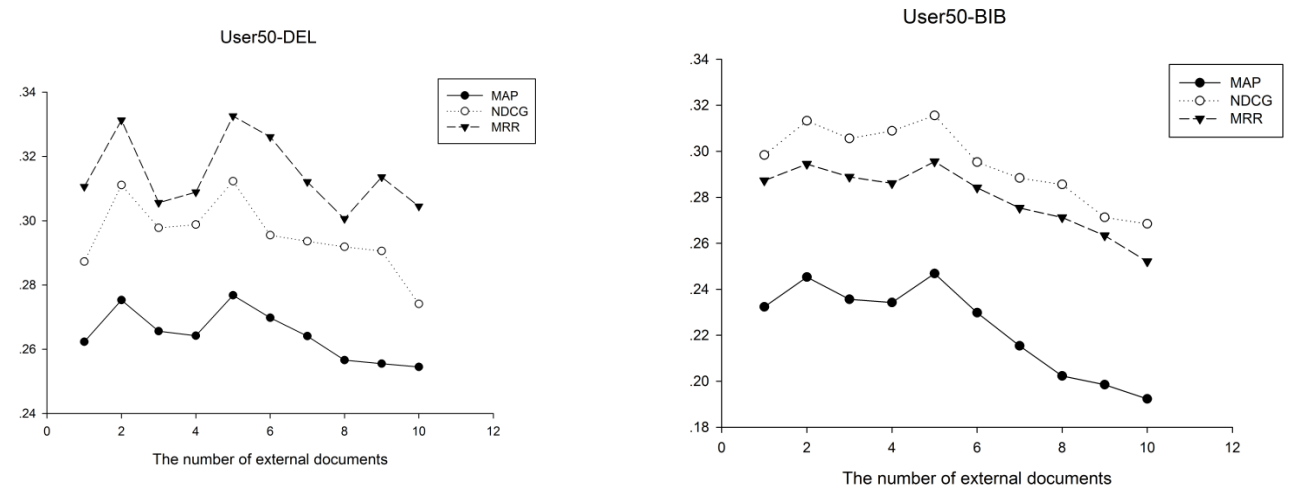

Fig. 5. The performance by varying the number of external documents on two corpora

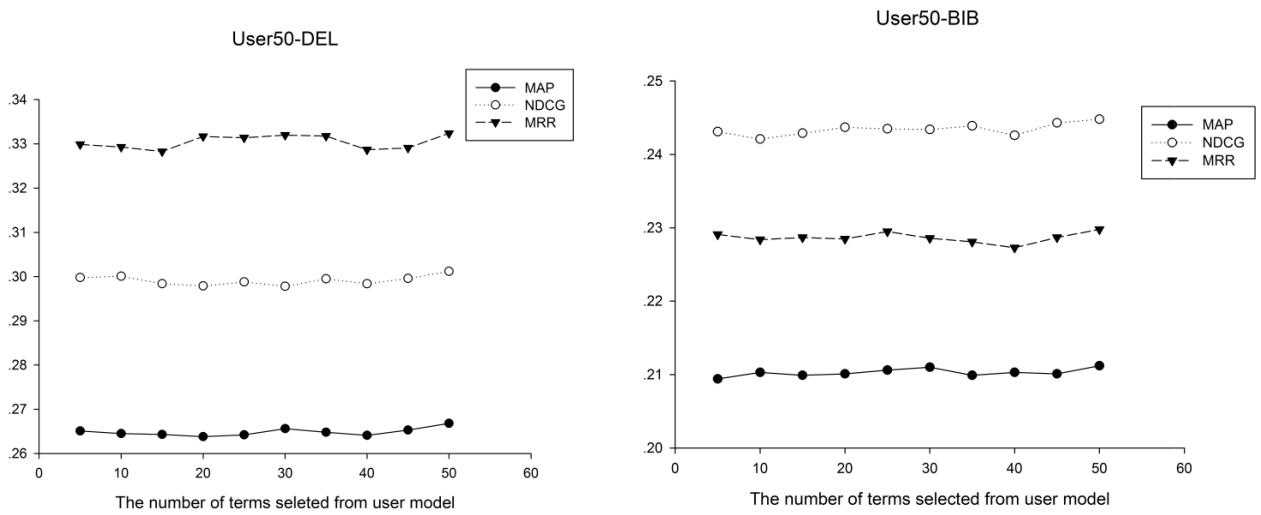

Fig. 6. The performance by varying the number of terms selected from user model on two corpora

our method increases from $k = 5$ to $k = 15$, while they are always fluctuated after $k = 15$. But when the number of topic is set to $k = 15$, our proposed method achieved the best performance by three evaluation metrics. The right picture of Fig.3 demonstrates that the optimal parameters of topic, which is consistent with the results on the DEL collection. The results show the number of topic can influence the performance in personalized search.

5.2.4 Impact of the Number of Word Embeddings

We also examine the performance about our method by the number of Word Embeddings. The process of examination is as the impact of the number of topics. The results are shown in Fig.4. As illustrated by the results, when the number of Word Embeddings is small, the increase of dimensions of Word Embeddings improves the performance of our method, since with the larger dimension of Word Embeddings, we can capture more context semantic information of the target word. However, when the dimension of Word Embeddings becomes too large, the performance of our method starts to decrease, this is because the dimension of Word Embeddings are set to large, which may bring too much noise and affect the accuracy of semantic information.

5.2.5 Impact of the Number of External Documents

The number of documents retrieved by each query (external documents) can significantly influence the performance of our proposed method in this paper. These documents fetched from external knowledge base, we called it as external documents. All the process of verification about it is still the same as the above section. It can be seen from Fig.5 that there is no obvious

difference between 2 and 5, but when the number of external document is set to 5, the performance achieved by our proposed method is the best. There is another obvious observation is that when the documents exceeds 5, the performance would be noticeable drop, which means too much noise introduced by too much documents. Thus the appropriate number of external documents is 5.

5.2.6 Impact of the Number of Expansion Terms

In this section, we describe the impact of the number of expansion terms, which is the number of terms selected from user model. From Fig.6, on the DEL collection and the BIB collection, in most of case, it is easy to tell us that there is generally consistent trend across all evaluated metrics. When the number of terms selected from user model is set to 50, it is clear that our method shows the best significance performance measured by three metrics. The reason is that a large number of expansion terms is selected from user model, which may be beneficial to choose more terms relevant to the original query.

6. Conclusions

We study the problem of personalized search utilizing social tagging data in the current paper. In particular, we investigated augmented user models and query expansion methods. We construct an augmented user model from a set of tags and documents, together with an external knowledge base. A novel generative model is proposed, which leverages word embeddings with Latent Dirichlet Allocation for user model generation. Based on the user models constructed, we further present a query expansion model to facilitate the desired personalized retrieval based on topics learnt. The proposed method performed well on two real-world social tagging datasets. It demonstrates statistically significant improvements over several baseline systems including non-personalized and personalized methods. In future research, automatically determination of the number of topics and dimensions will be studied. The effectiveness of different external knowledge bases will also be examined in our subsequent experiments.

Acknowledgments. The work described in this paper is supported by research was supported by Scientific Research Fund of Hunan Provincial Education Department of China Grant No. 16K030, Hunan Provincial Natural Science Foundation of China under No. 2017JJ2101, National Natural Science Foundation of China under Project No. 61300129, No. 61572187 and No. 61272063, and Hunan Provincial Innovation Foundation For Postgraduate (CX2016B575). This work is also supported by ADAPT Centre for Digital Content Technology, which is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2016) and is co-funded under the European Regional Development Fund.

References

[1] Ghorab, M.R., Zhou, D., O'Connor, A., Wade, V. (2013) Personalised Information Retrieval: survey and classification. User Modeling and User-Adapted Interaction 23, 381-443.

[2] Zhou, D., Lawless, S., Wu, X., Zhao, W., Liu, J. (2016) A study of user profile representation for personalized cross-language information retrieval. Aslib Journal of Information Management 68, 448-477.

[3] Xu, S., Bao, S., Fei, B., Su, Z., Yu, Y. (2008) Exploring folksonomy for personalized search. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 155-162.

[4] Bouadjenek, M.R., Hacid, H., Bouzeghoub, M. (2013) Sopra: a new social personalized ranking function for improving web search. Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp. 861-864.

[5] Xie, H., Li, X., Wang, T., Chen, L., Li, K., Wang, F.L., Cai, Y., Li, Q., Min, H. (2016) Personalized search for social media via dominating verbal context. Neurocomputing 172, 27-37.

[6] Xie, H., Li, X., Wang, T., Lau, R.Y.K., Wong, T.-L., Chen, L., Wang, F.L., Li, Q. (2016) Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy. Information Processing & Management 52, 61-72.

[7] Bouadjenek, M.R., Hacid, H., Bouzeghoub, M., Daigremont, J. (2011) Personalized social query expansion using social bookmarking systems. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 1113-1114.

[8] Zhou, D., Lawless, S., Wade, V.: (2012) Improving search via personalized query expansion using social media. Information Retrieval 15, 218-242.

[9] Zhou, D., Lawless, S., Wade, V. (2012) Web Search Personalization Using Social Data. Second International Conference on Theory and Practice of Digital Libraries:, TPDL 2012, pp. 298-310.

[10] Bender, M., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Parreira, J.X., Schenkel, R., Weikum, G. (2008) Exploiting social relations for query expansion and result ranking. Proceedings of the IEEE 24th International Conference on Data Engineering Workshop, ICDEW 2008, pp. 501-506. IEEE, USA.

[11] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. (2013) Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, NIPS 2013, pp. 3111-3119.

[12] Blei, D.M., Ng, A.Y., Jordan, M.I. (2003) Latent dirichlet allocation. Journal of Machine Learning Research. 3, 993-1022.

[13] Dou, Z., Song, R., Wen, J.-R. (2007) A large-scale evaluation and analysis of personalized search strategies. Proceedings of the 16th international conference on World Wide Web, pp. 581-590.
[14] Zhou, D., Lawless, S., Liu, J., Zhang, S., Xu, Y. (2015) Query expansion for personalized cross-language information retrieval. Proceedings of the 10th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2015, pp. 1-5.
[15] Chirita, P.-A., Firan, C.S., Nejdl, W. (2007) Personalized query expansion for the web. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 7-14.
[16] Wang, Q., Jin, H. (2010) Exploring online social activities for adaptive search personalization. Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 999-1008.
[17] Cai, Y., Li, Q. (2010) Personalized search by tag-based user profile and resource profile in collaborative tagging systems. Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 969-978.
[18] Das, R., Zaheer, M., Dyer, C. (2015) Gaussian LDA for Topic Models with Word Embeddings. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, pp. 795-804.
[19] Liu, Y., Liu, Z., Chua, T.-S., Sun, M. (2015) Topical Word Embeddings. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI 2015, pp. 2418-2424.
[20] Lavrenko, V., Croft, W.B. (2001) Relevance based language models. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 120-127.
[21] Ganguly, D., Leveling, J., Jones, G.F.: Topical Relevance Model. Information Retrieval Technology, vol. 7675, pp. 326-335. Springer Berlin Heidelberg (2012)
[22] Zubiaga, A., Garcia-Plaza, A.P., Fresno, V., Martinez, R. (2009) Content-based clustering for tag cloud visualization. Proceedings of the International Conference on Advances in Social Network Analysis and Mining, ASONAM'09, pp. 316-319.
[23] Zubiaga, A., Fresno, V., Martinez, R., Garcia-Plaza, A.P. (2013) Harnessing folksonomies to produce a social classification of resources. IEEE Transactions on Knowledge and Data Engineering, 25, 1801-1813.
[24] Zhai, C., Lafferty, J. (2001) Model-based feedback in the language modeling approach to information retrieval. Proceedings of the tenth international conference on Information and knowledge management, pp. 403-410.
[25] Diaz, F., Metzler, D. (2006) Improving the estimation of relevance models using large external corpora. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 154-161.