Enhanced Personalized Search using Social Data

Dong Zhou[1] **, Séamus Lawless**[2] **, Xuan Wu**[1] **, Wenyu Zhao**[1]**, Jianxun Liu**[1]
1. Key Laboratory of Knowledge Processing and Networked Manufacturing & School of Computer Science and Engineering, 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
dongzhou1979@hotmail.com, seamus.lawless@scss.tcd.ie

Abstract

Search personalization that considers the social dimension of the web has attracted a significant volume of research in recent years. A user profile is usually needed to represent a user's interests in order to tailor future searches. Previous research has typically constructed a profile solely from a user's usage information. When the user has only limited activities in the system, the effect of the user profile on search is also constrained. This research addresses the setting where a user has only a limited amount of usage information. We build enhanced user profiles from a set of annotations and resources that users have marked, together with an external knowledge base constructed according to usage histories. We present two probabilistic latent topic models to simultaneously incorporate social annotations, documents and the external knowledge base. Our web search strategy is achieved using personalized social query expansion. We introduce a topical query expansion model to enhance the search by utilizing individual user profiles. The proposed approaches have been intensively evaluated on a large public social annotation dataset. Results show that our models significantly outperformed existing personalized query expansion methods which use user profiles solely built from past usage information in personalized search.

1 Introduction

On today's social web, users can enrich the social context of web pages. The most notable fact is that users can often freely tag web pages with annotations (Gupta et al., 2011). These tags could be high quality descriptors of the web pages' topics and a good indicator of web users' interests. However, the uncontrolled manner of social tagging results in the use of an unrestricted vocabulary. This makes searching through the collection difficult and generally less accurate. Thus the social annotation or bookmarking system demonstrates an extreme example of the vocabulary mismatch problem encountered in personalized web search. To tackle the problem, various personalized query expansion (QE) and results reranking techniques have been proposed and evaluated (Bouadjenek et al., 2016).

There have been some attempts to achieve personalized QE using social data. For example, Researchers have considered selecting the most related tags from a user's profile to expand queries (Bender et al., 2008; Bertier et al., 2009; Bouadjenek et al., 2011). Local analysis and co-occurrence based user profile representation have also been adopted to expand the query (Chirita et al., 2007; Biancalana et al., 2013). Recently, Zhou et al. proposed a query expansion framework based on individual user profiles (Zhou et al., 2012a). In their work, terms in the user profile are modeled according to their associations, which can be defined by co-occurrence statistics or defined by a tag-topic model.

All of the previously mentioned systems are dependent upon historical usage information being available in an individual user profile (Sugiyama et al., 2004; Teevan et al., 2005; Bennett et al., 2012; Zhou et al., 2014; Guha et al., 2015; Zhou et al., 2016). This information is pivotal when tailoring search results to the preferences of specific individuals. However, in some

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 700–710,
Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics

cases a user may have very limited previous interactions with the system. With little usage information to hand, the personalized search experience is poor. Furthermore, using only historical usage information to personalize search may not be enough.

In this paper, we extend personalized search using social data in two directions. First, we exploit external knowledge bases to enhance the user profile built from a user's historical usage information. We build queries from the user tags and annotated web pages. Subsequently, we fetch the relevant documents from an external corpus to be included in the user profile. We then propose to incorporate the user's annotations, web page content information and external documents through two statistical models, which we have named Mixture Enhanced User Profiling (MEUP) Model and Separated Enhanced User Profiling (SEUP) Model. Both models infer latent topics, their probabilities of being relevant and a multinomial distribution of topics of the documents being considered. MEUP mixed the tags, annotated documents and external documents together to infer unified latent topics, while SEUP is an extension of MEUP which learns topics that are shared between the two groups of document-aligned pairs.

Second, we propose a topical query expansion model to personalize web search by utilizing the user profiles. In the topical QE model, profile terms are calculated based on their topical relevance to the query terms to expand the query.

Experimental results show that the Enhanced User Profiling models together with the topic QE can significantly improve retrieval performance over user profiles solely built from a user's historical information. Improvements were observed for users with both a rich amount of usage information and a small amount of information. We also demonstrate that the approach proposed in the paper outperforms existing QE methods proposed for personalized search using social data.

The contribution of this paper can be summarized as follows:

i. *We tackle the challenge of personalized web search using social data in a novel way by enhancing user profiles that are built solely from users' historical usage information.*

ii. *We propose and systematically evaluate two novel generative models to construct enriched user profiles with the help of external corpora in the context of personalized search using social data.*

iii. *We suggest and evaluate a novel query expansion method. Instead of relying on lexical relevance information between query terms and profile terms, we also consider the topical relevance between them to expand the query.*

2 Related Work

2.1 Personalized Search Using Social Media

In personalized search using social media (Jamali and Ester, 2010; Lin et al., 2013), the search process is either performed over "social" data gathered from Web 2.0 applications such as social bookmarking systems, wikis, blogs etc., or it re-adapts the web search results produced by search engines by using social data (Carman et al., 2008; Bouadjenek et al., 2016). For example, the authors in (Vallet et al., 2010) investigated how the ranking of search engine results can be improved with respect to users if the users' social information is taken into consideration. A similar approach was also explored in (Noll and Meinel, 2007) where the system performed re-ranking of Google search results based on social bookmarks and tags harvested from *del.icio.us*[1]. 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.

2.2 Personalized Results Re-Ranking

Because of this problem, researchers started to use social data as a test collection to develop personalized techniques. In this way, personalization usually involves two general approaches. The first approach submits a query into the collection but re-ranks the returned results based on an individual user profile. In (Xu et al., 2008) the authors re-rank the results according to the topical relevance of documents and users' interests. Carmel et al. (Carmel et al., 2009) investigated personalized results re-ranking based on the user's social relations. Wang and Jin (Wang and Jin, 2010) explored gathering data from multiple online social systems for adaptive search personalization. Bouadjenek et al. (Bouadjenek et al., 2013a; Bouadjenek et al., 2013b) propose to use social data and user relationships to enhance document

[1] http://www.delicious.com

Notation	Meaning
$\mathcal{U}$	finite sets of users
$\mathcal{D}$	finite sets of web pages/documents
$\mathcal{T}$	finite sets of tags
$\mathcal{A}$	a ternary relation, elements are tags
$\mathcal{A}^u$	the set of annotations of a user
$\mathcal{T}^u$	the tag vocabulary of a user
$\mathcal{D}^u$	a user's set of documents
t	a tag
d	a document
u	a user
w	a word/term
$docTerm^u$	the vocabulary extracted from the documents that a user has tagged
$exterTerm^u$	the full set of terms extracted from a user's external documents
$\mathcal{D}_{exter}$	an external corpus
$\mathcal{D}^u_{exter}$	a user's set of external documents
q	a source query
$q^{\mathcal{T}^u}$	a query concatenated by tags of a user
$Q^{\mathcal{D}^u}$	queries extracted from a user's set of documents
Q_{exter}	queries to be sent to an external corpus
$s_{x,y}$	retrieval score of a query q_y to retrieve a document d_x
K	number of topics
μ_z	mean of Log-normal distribution of retrieval scores for topic z
σ_z	deviation of Log-normal distribution of retrieval scores for topic z
N_{d_j}	number of `words` in document d_j
$z_{j,i}$	topic associated with the i-th word in the document d_j
$w_{j,i}$	i-th word in document d_j
$n_{j,k}$	the number of times that topic k sampled w.r.t. document d_j
$v_{k,w_{j,i}}$	the number of times $w_{j,i}$ has been generated by topic k
θ	multinomial distribution of topics
φ	multinomial distribution of words
ϕ	multinomial distribution of words (used in SEUP)
α	the parameter of topic Dirichlet prior
β	the parameter of word Dirichlet prior

Table 1. Basic notations used in the paper

representation for re-ranking purposes. Though this group of work is attractive, if relevant items cannot be fetched in the first place, regardless of the complex re-ranking process, the results still tend to be unsatisfactory.

2.3 Personalized Query Expansion

Another group of work modifies or augments a user's original query. This approach is termed query expansion (Zhou et al., 2015). Researchers have considered tag-tag relationships for personalized query expansion, by selecting the most related tags from a user's profile (Bender et al., 2008; Bertier et al., 2009). However, tags cannot be relied upon to consistently provide precise descriptions of resources for use when searching. Local analysis and co-occurrence based user profile representation have also been adopted to expand the query (Chirita et al., 2007; Biancalana and Micarelli, 2009). However, the expansion terms are solely based on lexical matching between the query and the terms which exist in the user profile. Zhou et al. proposed a query expansion framework based on individual user profiles (Zhou et al., 2012a; Zhou et al., 2012b). In their work, terms in the user profile are modeled according to their associations, which can be defined by co-occurrence statistics or defined by a tag-topic model. The method simultaneously incorporates annotations and web documents in a latent graph, regularized by terms extracted from the top-ranked documents.

However, all of the previously mentioned systems consider constructing user profiles solely from past usage information. In contrast, in this paper we extend personalized web search using social data by exploiting an external knowledge base to enhance the user profile.

3 Problem Definition and Solution Overview

In social annotation and bookmarking systems such as *del.icio.us* or *CiteUlike*[2], 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* website. Textual content can be crawled by following a URL that refers to a *document* or *web page*. Please refer to Table 1 for the basic notations used in this paper.

Formally, social tagging data 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 (or bookmarks). The

[2] http://www.citeulike.org

Input: A query q
Tags of a user $\mathcal{T}^u$
Documents of a user $\mathcal{D}^u$
An external corpus $\mathcal{D}_{exter}$
Output: An expanded query q'

/* step one: External documents fetch */
1. $q^{\mathcal{T}^u} \leftarrow \cup(t \in \mathcal{T}^u)$
2. **for all** $d_x \in \mathcal{D}^u$ **do**
3. $Q^{\mathcal{D}^u} \leftarrow EXTRACTTOP(w \in d_x)$
4. $Q_{exter} \leftarrow q^{\mathcal{T}^u} \cup Q^{\mathcal{D}^u}$
5. **for all** $q_y \in Q_{exter}$ **do**
6. $\mathcal{D}^u_{exter} \leftarrow RETRIEVE_{\mathcal{D}_{exter}}(q_y)$
7. Record retrieval score $s_{x,y}$
/* step two: User profile modelling */
8. **for** $k \in [1, K]$ **do**
9. Initialize μ_z and σ_z randomly
10. **for** $d_j \in \mathcal{T}^u \cup \mathcal{D}^u \cup \mathcal{D}^u_{exter}$ **do**
11. **for** w_i indexed by $i = 1, \dots, N_{d_j}$ **do**
12. Draw $z_{j,i}$ from $p(z_{j,i} = k)$
13. Update $n_{j,k}$ and $v_{k,w_{j,i}}$
14. Calculate the posterior estimate of θ and φ
/* step three: Personalized query expansion */
15. $\{w_1, w_2 \dots w_n\} \leftarrow exterTerm^u \cup docTerm^u \cup \mathcal{T}^u$
16. **for all** $w \in \{w_1, w_2 \dots w_n\}$ **do**
17. calculate $P(w|q)$ using topics from step two
18. Output q' consists of top δ terms with the highest $P(w|q)$

Table 2. Procedure for personalized search using social data

set of annotations of a user is defined as: $\mathcal{A}^u \coloneqq \{(t, d)|u, d, t \in \mathcal{A}\}$. The tag vocabulary of a user, is given as $\mathcal{T}^u \coloneqq \{t|(t, d) \in \mathcal{A}^u\}$. A user's set of documents is $\mathcal{D}^u \coloneqq \{d|(t, d) \in \mathcal{A}^u\}$. We define terms extracted from a user's set of documents as $docTerm^u \coloneqq \{w|w \in \mathcal{D}^u\}$, where w denotes a word/term in the annotated documents. Similarly, we define terms extracted from a user's set of external documents as $exterTerm^u \coloneqq \{w|w \in \mathcal{D}^u_{exter}\}$, 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 profile $\{w_1, w_2 \dots w_n\}$ the goal is to return a ranked list of profile terms to be added to the query, for a second round retrieval of results.

Our personalization approach consists of three main steps (see Table 2): Fetching external documents; User profile modelling; and Personalized query expansion. We enhance a user's historical usage information in step one. We firstly concatenate all tags t in $\mathcal{T}^u$ into a query $q^{\mathcal{T}^u}$ (line 1). Then for each document d in $\mathcal{D}^u$, we extract terms with the highest inverted document frequency (*idf*) scores as queries $Q^{\mathcal{D}^u}$ (lines 2-3, with the *EXTRACTTOP* function returns top λ terms). Next we send queries in Q_{exter} ($q^{\mathcal{T}^u} \cup Q^{\mathcal{D}^u}$) to an external corpus $\mathcal{D}_{exter}$ to fetch $\mathcal{D}^u_{exter}$ together with their retrieval scores $s_{x,y}$ (lines 5-7, the number of documents retrieved by each query is controlled by the parameter γ). Step two integrates $\mathcal{T}^u$ (here all tags are concatenated and viewed as a single document), $\mathcal{D}^u$, $\mathcal{D}^u_{exter}$ and their retrieval scores $s_{x,y}$ into a topic model such that a multinomial distribution of topics specific to each document can be inferred (lines 8-14, we eliminate the procedure for the SEUP model because its similarity to the simpler model, see the next section). In the last step, the algorithm uses the output of step two to build a topical query expansion model in order to expand the original query (lines 15-18). Note that step one and step two could be executed off-line so as to improve the efficiency of the algorithm.

4 Enhanced User Profiling Models

In this section we describe how to model user profiles (i.e. step two in Table 2). We present two Enhanced User Profiling (EUP) models for this purpose.

4.1 Mixture Enhanced User Profiling

Topic discovery in the EUP models is influenced not only by term co-occurrences, but also by the retrieval scores of documents. To avoid normalization, we employ a log-normal distribution for retrieval scores to infer latent topics via the documents and their relevance probabilities.

The MEUP model developed here is a generative model of retrieval scores and the words in the documents. The generative process is as follows:

Generative process of the MEUP model

1. **for** each topic $k \in [1, K]$ **do**
sample the mixture of words $\varphi \sim Dirichlet(\beta)$
2. **for** each document $d_j \in \mathcal{T}^u \cup \mathcal{D}^u \cup \mathcal{D}^u_{exter}$ **do**
sample the mixture of topics $\theta_j \sim Dirichlet(\alpha)$
for each word w_i indexed by $i = 1, \dots, N_{d_j}$ **do**
sample the topic index topic $z_{j,i} \sim Mult(\theta_{d_j})$
sample the weight of word $w_{j,i} \sim Mult(\varphi_{z_{j,i}})$
sample the retrieval score $s_{,i} \sim \mathcal{N}(\mu_{z_{j,i}}, \sigma_{z_{j,i}})$

In the above process, the retrieval scores of terms in the same document are the same and calculated by a language model retrieval function (Manning et al., 2008) for retrieved documents in $\mathcal{D}^u_{exter}$. The retrieval scores for $\mathcal{T}^u$ (here all tags are concatenated and viewed as a single document) and documents in $\mathcal{D}^u$ are set to one. We normalize the scores by the max score in the retrieval list. We used a fixed number of latent topics K. The posterior distribution of topics depends on two sets of information, both the terms and retrieval scores of the documents.

In this model, inference is intractable. We use Gibbs Sampling (Heck and Thomas, 2015) to perform approximate inference. We employ a conjugate prior for the multinomial distributions, and integrate out θ and φ. In the sampling procedure, we need to calculate the conditional distribution $p(z_{j,i} = k)$ (line 12 in Table 2). By using Gibbs Sampling, for each word the topic is sampled from:

$$p(z_{j,i} = k) \propto \frac{n_{j,k,\neg i} + \alpha}{n_{j,\cdot,\neg i} + K \cdot \alpha} \times \frac{v_{k,w_{j,i},\neg} + \beta}{v_{k,\cdot,\neg} + V \cdot \beta} \times \frac{1}{s_{j,i}\sigma_{z_{j,i}}\sqrt{2\pi}} \exp\left(-\frac{\left(\ln s_{j,i} - \mu_{z_{j,i}}\right)^2}{2\sigma^2_{z_{j,i}}}\right)$$

where $n_{j,k,\neg i}$ counts the number of times that topic with index k has been sampled from the multinomial distribution specific to document d_j with the current $z_{j,i}$ not counted. Another counter variable $v_{k,w_{j,i},\neg}$ counts the number of times $w_{j,i}$ has been generated by topic k, but not counting the current $w_{j,i}$. A dot denotes summation over all values of the variable whose index that dot takes. $\mu_{z_{j,i}}$ and $\sigma_{z_{j,i}}$ are elements from μ_z and σ_z, respectively. After that we can calculate the posterior estimate of θ and φ (line 14 in Table 2).

4.2 Separated Enhanced User Profiling

In the MEUP model, $\mathcal{T}^u$, $\mathcal{D}^u$ and $\mathcal{D}^u_{exter}$ are mixed together to infer unified latent topics. However, the MEUP model may miss important information when the topics are learned. Our SEUP model extends the MEUP model by learning 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 step one of our approach, we treat the document (from $\mathcal{D}^u_{exter}$) and the query (from Q_{exter}) as pseudo-aligned documents in two groups. The first group we named source group C, the other group we named target group E. By using the aligned documents, we propose a model to learn the latent topics between the two groups.

Note that in this case, there is a comparable document set aligned at the document-level. Therefore, θ can be viewed as a group independent factor, and shared among comparable aligned documents. Henceforth, the generation process for the SEUP model is slightly different from the MEUP model. The generative process is summarized below:

Generative process of the SEUP model

1. **for** each of topics $k \in [1, K]$ **do**
   sample the mixture of words $\varphi \sim Dirichlet(\beta)$
   sample the mixture of words $\phi \sim Dirichlet(\beta)$
2. **for** each document pair $d_j = \{d^C_j \in \mathcal{T}^u \cup \mathcal{D}^u, d^E_j \in \mathcal{D}^u_{exter}\}$ **do**
   sample the mixture of topics $\theta_j \sim Dirichlet(\alpha)$
   for each word w^C_i indexed by $i = 1, \ldots, N_{d^C_j}$ **do**
   sample the topic index topic $z^C_{j,i} \sim Mult(\theta_{d_j})$
   sample the weight of word $w^C_{j,i} \sim Mult(\varphi_{z^C_{j,i}})$
   sample the retrieval score $s^C_{j,i} \sim \mathcal{N}(\mu^C_{z_{j,i}}, \sigma^C_{z_{j,i}})$
   for each word w^E_i indexed by $i = 1, \ldots, N_{d^E_j}$ **do**
   sample the topic index topic $z^E_{j,i} \sim Mult\left(\theta_{d_j}\right)$
   sample the weight of word $w^E_{j,i} \sim Mult(\phi_{z^E_{j,i}})$
   sample the retrieval score $s^E_{j,i} \sim \mathcal{N}(\mu^E_{z_{j,i}}, \sigma^E_{z_{j,i}})$

Similar to the MEUP model, the updated formulas for Gibbs sampling for the SEUP model are:

$$p(z^C_{j,i} = k) \propto \frac{n^C_{j,k,\neg i} + n^E_{j,k} + \alpha}{n^C_{j,\cdot,\neg i} + n^E_{j,\cdot} + K \cdot \alpha} \times \frac{v^C_{k,w^C_{j,i},\neg} + \beta}{v^C_{k,\cdot,\neg} + V^C \cdot \beta} \times \frac{1}{s^C_{j,i}\sigma^C_{z_{j,i}}\sqrt{2\pi}} \exp\left(-\frac{\left(\ln s^C_{j,i} - \mu^C_{z_{j,i}}\right)^2}{2\sigma^{C\,2}_{z_{j,i}}}\right)$$

$$p(z^E_{j,i} = k) \propto \frac{n^E_{j,k,\neg i} + n^C_{j,k} + \alpha}{n^E_{j,\cdot,\neg i} + n^C_{j,\cdot} + K \cdot \alpha} \times \frac{v^E_{k,w^E_{j,i},\neg} + \beta}{v^E_{k,\cdot,\neg} + V^E \cdot \beta} \times \frac{1}{s^E_{j,i}\sigma^E_{z_{j,i}}\sqrt{2\pi}} \exp\left(-\frac{\left(\ln s^E_{j,i} - \mu^E_{z_{j,i}}\right)^2}{2\sigma^{E\,2}_{z_{j,i}}}\right)$$

The meaning of the symbols used in the SEUP model is the same as in the MEUP model, except this time for two groups *E* and *C*. In the two EUP models, the multinomial distribution of topics is specific to each document and each word can be easily inferred.

5 Topical Query Expansion

In step three of our approach to personalization, we use the output from step two to build a QE model that calculates the weights of the profile terms to be added to the initial query. In this section we detail this process.

Given the query $q = \{w_a^q\}_{a=1}^n$ of n independent query terms, the probability of the query generating a word w is defined as (see also (Lavrenko and Croft, 2001; Ganguly et al., 2012)):

$$P(w|q) = P(w|w_1^q, \dots, w_n^q) \propto \prod_{a=1}^{n} P(w|w_a^q)$$

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:

$$P(w|w_a^q) = \sum_{b=1}^{N} P(w|d_b)P(d_b|w_a^q) \propto \frac{1}{N}\sum_{b=1}^{N} P(w|d_b)P(w_a^q|d_b)$$

The calculation discards the uniform prior for $P(w_a^q)$, and takes the uniform prior of documents outside the summation.

As we already have outputs from step two, the documents inside the user profile can be used as a set of relevant documents in the above calculation. In addition, because we now have latent topics related to each document and each word, there is no longer a direct dependency of w on d_b and q. 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:

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

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

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

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

$$P(w|w_a^q) \propto \frac{1}{N}\sum_{b=1}^{N}(\sum_{k=1}^{K} P(w|z_k)P(z_k|d_b))\times(\sum_{k=1}^{K} P(w_a^q|z_k)P(z_k|d_b))$$

In the SEUP 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 $\{w_1, w_2 \dots w_n\} = exterTerm^u \cup docTerm^u \cup \mathcal{T}^u$ are ranked by their probability of being generated by the given query $P(w|q)$ (line 16-17 in Table 2), and the top δ terms are chosen to expand the query.

6 Evaluation

In the following section we describe experiments which have been designed to evaluate the proposed method. We start the section by discussing the experimental settings, and then we present and analyze the results.

6.1 Experimental Setup

In order to evaluate the above proposed methods on real-world data, we selected two delicious datasets: *socialbm0311* and *deliciousT140*, which are public, described and analyzed in (Zubiaga et al., 2009; Zubiaga et al., 2013). The *deliciousT140* dataset is made up by 144,574 unique URLs, all of them with their corresponding social tags retrieved from *del.icio.us*. However, this dataset does not contain the actual web pages (i.e. documents). So we used another *socialbm0311* dataset. It 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. We used a public parser[3] to parse the web pages in order to get their textual content.

We constructed two corpora from different external knowledge bases. The first corpus was obtained from the largest encyclopedia – Wikipedia[4]. 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)[5], which we refer to as CLEF. This collection contains 304,630 documents.

To investigate the effects of enhanced user profiles, we selected two groups of users as test users. One group contains 1,000 randomly selected users with no more than 50 bookmarks (refer to as **User-SMALL**) and another group contains 1,000 randomly selected users with more than 200 bookmarks (refer to as **User-LARGE**). These two groups of users represent users with small amount

[3] http://htmlparser.sourceforge.net/
[4] http://www.wikipedia.org
[5] http://catalog.elra.info/

of and rich amount of past usage information respectively. The English terms were processed by down-casing the alphabetic characters, removing the stop words and stemming words using the Por ter stemmer. For each user, 75% of his/her tags with annotated web pages were used to create the user profile and the other 25% were used as a test collection.

The evaluation method used by previous researchers in personalized social search (Xu et al., 2008; Wang and Jin, 2010; Zhou et al., 2012a) 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).

The following evaluation metrics were chosen to measure the effectiveness of the various approaches: the normalized discounted cumulative gain (NDCG), mean reciprocal rank (MRR) and mean average precision (MAP) (Voorhees, 1999; Järvelin and Kekäläinen, 2000). The average performance over all users is calculated. Statistically significant differences were determined using a paired t-test at a confidence level of 95%.

6.2 Experimental Runs

The proposed approach is applied to social search personalization through the means of query expansion. We evaluate our proposed models and compare with several state-of-the-art methods as follows.

LM A popular and quite robust language model retrieval method which has previously demonstrated good results (Zhai and Lafferty, 2001). We compute the Kullback-Leibler divergence between the query and document language model as described in (Zhai and Lafferty, 2001).

LMRM A relevance model involves pseudo-relevance feedback in the language model as in (Lavrenko and Croft, 2001). We include this model as a competitive non-personalized query expansion baseline.

LMRM-external This is a modified version of the relevance model as described in (Diaz and Metzler, 2006). 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 corpus and CLEF.

Co-occur This method has been used by several researchers. In this method the selection of expansion terms is based on their co-occurrence statistics with the query terms and other terms inside the user model. We used this approach as previously it demonstrated satisfactory performance as in (Chirita et al., 2007).

Co-tag Pure tag-tag relationships are also favored by many researchers. This method is based on the co-tagging activities a user performed (Bender et al., 2008; Bertier et al., 2009; Bouadjenek et al., 2011). In this case, the user profiles contain training tags with their co-tagging statistics computed using the Jaccard coefficient.

Tag-topic-regu Zhou et al. (Zhou et al., 2012a) proposed a query expansion framework based on regularizing the smoothness of word associations over a connected graph using terms extracted from top-ranked documents. The user profiles are built according to a Tag-Topic model in a latent graph. We include the highest performing method from their work for comparison.

MEUP From our proposed methods, the MEUP method using the MEUP model and the topical query expansion method for social web search.

SEUP This is our alternative proposed method, by using the SEUP model and the topical query expansion method to personalize search.

The number of documents retrieved by each query in step one is set to $\gamma = 5$ empirically. Parameter λ used in the *EXTRACTTOP* function is set to 10. For the EUP 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 5. All the parameters in the other baseline models are set according to their tuning procedures in the original papers

6.3 Results

Firstly we examine 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, which are shown in Table 3. The statistically significant differences are marked as *l* and *w* with respect to the **LMRM** and **LMRM-external** baselines as these two methods work better than the simpler **LM**

User-SMALL

	Wikipedia			CLEF		
	MAP	**NDCG**	**MRR**	**MAP**	**NDCG**	**MRR**
LM	0.0216	0.0449	0.0226	0.0216	0.0449	0.0226
LMRM	0.0241	0.0547	0.0261	0.0241	0.0547	0.0261
LMRM-external	0.0283	0.0588	0.0307	0.0272	0.0585	0.0290
Co-occur	0.0499 $^{l, w}$	0.0812 $^{l, w}$	0.0600 $^{l, w}$	0.0499 $^{l, w}$	0.0812 $^{l, w}$	0.0600 $^{l, w}$
Co-tag	0.0491 $^{l, w}$	0.0758 $^{l, w}$	0.0538 $^{l, w}$	0.0491 $^{l, w}$	0.0758 $^{l, w}$	0.0538 $^{l, w}$
Tag-topic-regu	0.0597 $^{l, w, o, t,}$	0.0955 $^{l, w, o, t,}$	0.0666 $^{l, w, o, t,}$	0.0597 $^{l, w, o, t,}$	0.0955 $^{l, w, o, t,}$	0.0666 $^{l, w, o, t,}$
MEUP	0.0729 $^{l, w, o, t, r}$	0.1058 $^{l, w, o, t, r}$	0.0844 $^{l, w, o, t, r}$	0.0722 $^{l, w, o, t, r}$	0.0981 $^{l, w, o, t, r}$	0.0770 $^{l, w, o, t, r}$
SEUP	0.0906 $^{l, w, o, t, r}$	0.1316 $^{l, w, o, t, r}$	0.1032 $^{l, w, o, t, r}$	0.0802 $^{l, w, o, t, r}$	0.1221 $^{l, w, o, t, r}$	0.0940 $^{l, w, o, t, r}$

User-LARGE

	Wikipedia			CLEF		
	MAP	**NDCG**	**MRR**	**MAP**	**NDCG**	**MRR**
LM	0.0178	0.0366	0.0194	0.0178	0.0366	0.0194
LMRM	0.0185	0.0400	0.0201	0.0185	0.0400	0.0201
LMRM-external	0.0195	0.0433	0.0263	0.0190	0.0420	0.0245
Co-occur	0.0386 $^{l, w}$	0.0578 $^{l, w}$	0.0409 $^{l, w}$	0.0386 $^{l, w}$	0.0578 $^{l, w}$	0.0409 $^{l, w}$
Co-tag	0.0381 $^{l, w}$	0.0546 $^{l, w}$	0.0399 $^{l, w}$	0.0381 $^{l, w}$	0.0546 $^{l, w}$	0.0399 $^{l, w}$
Tag-topic-regu	0.0470 $^{l, w, o, t,}$	0.0778 $^{l, w, o, t,}$	0.0498 $^{l, w, o, t,}$	0.0470 $^{l, w, o, t,}$	0.0778 $^{l, w, o, t,}$	0.0498 $^{l, w, o, t,}$
MEUP	0.0579 $^{l, w, o, t, r}$	0.0971 $^{l, w, o, t, r}$	0.0629 $^{l, w, o, t, r}$	0.0545 $^{l, w, o, t, r}$	0.0805 $^{l, w, o, t, r}$	0.0581 $^{l, w, o, t, r}$
SEUP	0.0633 $^{l, w, o, t, r}$	0.1049 $^{l, w, o, t, r}$	0.0678 $^{l, w, o, t, r}$	0.0604 $^{l, w, o, t, r}$	0.0978 $^{l, w, o, t, r}$	0.0651 $^{l, w, o, t, r}$

Table 3. Overall results, statistically significant differences between our methods and LMRM, LMRM-External, Co-occur, Co-tag, Tag-topic-regu are indicated by *l, w, o, t, r* respectively.

method. As illustrated by the results, the **LM** model was the lowest performer for all evaluation metrics for two groups of users. This result shows that merely borrowing common lexical-matching techniques from traditional information retrieval will not solve the personalized search problem. With the help of pseudo-relevance feedback, the **LMRM** and **LMRM-external** methods work consistently better than the **LM** baseline. This demonstrates the power of query expansion. Specifically, the technique that explores external corpora to obtain the relevant documents works better than the method which simply uses top-ranked documents. The results are consistent with previous research (Diaz and Metzler, 2006). The improvements are more noticeable when using Wikipedia as the external corpus. However, all the non-personalized baselines are outperformed by the personalized approaches including our proposed methods **MEUP** and **SEUP**, all with statistically significant results. This illustrates that non-personalized query expansion methods can only bring limited improvements while methods with additional terms from the user profiles can greatly improve retrieval effectiveness.

Next we evaluate the performance of the proposed methods compared to several personalized baselines that use only the users' past information for query expansion, i.e. **Co-occur**, **Co-tag** and **Tag-topic-regu** methods.

As seen from Table 3, three conclusions emerge. First, **MEUP** and **SEUP** both outperform all personalization methods previously proposed, in all metrics measured with two external corpora for both groups of 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 fact that our methods do not only consider a user's past usage information, but also uses an external knowledge base to enhance the user profiling process. Secondly, the **SEUP** method works consistently better than the **MEUP** method. This result confirms that merely mixing the documents from the historical evidence and external knowledge bases will miss some important information. By treating the documents as a pseudo-aligned corpus, we obtain much better performance. The highest improvement over the best performing run reaches 54.95% (in terms of the **SEUP** method with the MRR metric when compared to **Tag-topic-regu** by using Wikipedia as the external corpus in the **User-SMALL** group). Third, further improvements are achieved by us-

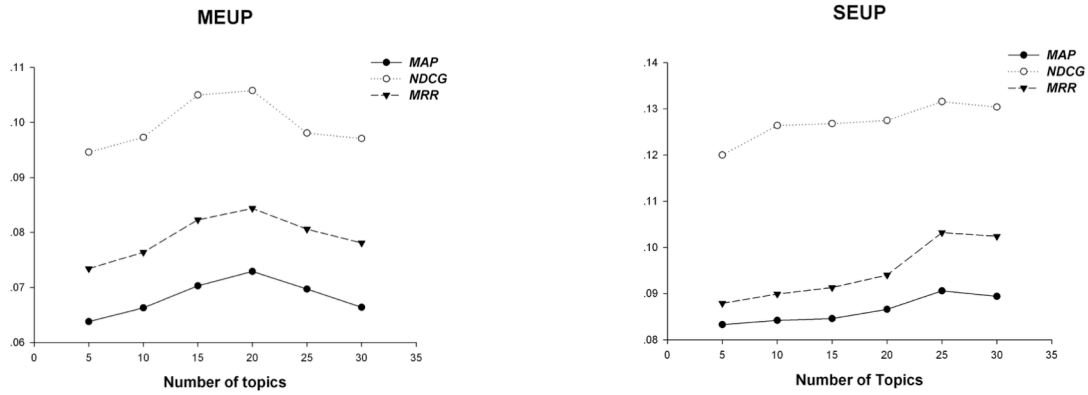

Fig. 1. Performance with different number of topics by using Wikipedia as external corpus in the user-SMALL group

ing Wikipedia as the external corpus rather than using the CLEF collection. The possible reason, as pointed out by Diaz and Metzler (Diaz and Metzler, 2006), is that an external corpus is likely to be a better source of expansion terms if it has better topic coverage over the target corpus. The results also show that the improvements over baseline models in the **User-SMALL** group are more noticeable than in the **User-LARGE** group. However, the differences are small. This result confirms that our methods work well both for users with small amounts, and those with rich amounts of past usage information.

We now examine the effect of the performance of the number of latent topics used in **MEUP** and **SEUP**. We vary the number of topics in both methods from 5 to 30, the results are shown in Figure 1, using Wikipedia as the external corpus in the **user-SMALL** group (we eliminated other results as they gave similar results). As can be seen from the figure, the highest performance is reached when the number of latent topics is 20 in **MEUP** and 25 in **SEUP**. When the number of topics continues to grow, the performance starts to degrade. However, even the lowest scored run still outperformed the strongest baseline. By varying the topic numbers, **SEUP** still outperforms **MEUP**.

7 Conclusion and Future Work

In this paper, we tackle the challenge of personalized web search using social data in a novel way by building enhanced user profiles from the annotations and resources the user has marked, together with an external knowledge base. We present two probabilistic latent models to simultaneously incorporate social annotations, documents and the external knowledge base. In addition, we introduce a topical query expansion model to enhance the search by utilizing individual user profiles. The proposed methods performed well on the social data crawled from the web, delivering statistically significant improvements over non-personalized and personalized representative baseline systems by constructing user profiles from a user's historical usage information only. It is also confirmed that our proposed methods work well for both active and less active users. In future research, we aim to automatically estimate the number of topics to be used in the EUP models. We also plan to explore the use of more external resources and novel latent semantic models to enhance performance.

Acknowledgements

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, Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, China under grant No. [2013] 1792, Hunan Provincial Innovation Foundation For Postgraduate under grant No. CX2016B575. This work is also supported by the ADAPT Centre for Digital Content Technology, which is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

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

Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk and Xiaoyuan Cui (2012). Modeling the impact of short- and long-term behavior on search personalization. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, Portland, Oregon, USA, ACM. p. 185-194.

Marin Bertier, Rachid Guerraoui, Vincent Leroy and Anne-Marie Kermarrec (2009). Toward personalized query expansion. In *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*, Nuremberg, Germany, ACM. p. 7-12.

Claudio Biancalana, Fabio Gasparetti, Alessandro Micarelli and Giuseppe Sansonetti (2013). Social semantic query expansion. *ACM Transactions on Intelligent Systems and Technology,* **4**(4): 1-43.

Claudio Biancalana and Alessandro Micarelli (2009). Social Tagging in Query Expansion: A New Way for Personalized Web Search. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, IEEE. p. 1060-1065.

Mohamed Reda Bouadjenek, Hakim Hacid and Mokrane Bouzeghoub (2013a). Sopra: a new social personalized ranking function for improving web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, Dublin, Ireland, ACM. p. 861-864.

Mohamed Reda Bouadjenek, Hakim Hacid and Mokrane Bouzeghoub (2016). Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms. *Information Systems* **56**: 1-18.

Mohamed Reda Bouadjenek, Hakim Hacid, Mokrane Bouzeghoub and Johann Daigremont (2011). Personalized social query expansion using social bookmarking systems. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, Beijing, China, ACM. p. 1113-1114.

Mohamed Reda Bouadjenek, Hakim Hacid, Mokrane Bouzeghoub and Athena Vakali (2013b). Using social annotations to enhance document representation for personalized search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, Dublin, Ireland, ACM. p. 1049-1052.

Mark J. Carman, Mark Baillie and Fabio Crestani (2008). Tag data and personalized information retrieval. In *Proceeding of the 2008 ACM workshop on Search in social media*, Napa Valley, California, USA, ACM. p. 27-34.

David Carmel, Naama Zwerdling, Ido Guy, Shila Ofek-Koifman, Nadav Har'el, Inbal Ronen, Erel Uziel, Sivan Yogev and Sergey Chernov (2009). Personalized social search based on the user's social network. In *Proceeding of the 18th ACM conference on Information and knowledge management*, Hong Kong, China, ACM. p. 1227-1236.

Paul - Alexandru Chirita, Claudiu S. Firan and Wolfgang Nejdl (2007). Personalized query expansion for the web. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, Amsterdam, The Netherlands, ACM. p. 7-14.

Fernando Diaz and Donald Metzler (2006). Improving the estimation of relevance models using large external corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, Seattle, Washington, USA, ACM. p. 154-161.

Debasis Ganguly, Johannes Leveling and GarethJ F. Jones (2012). Topical Relevance Model. *Information Retrieval Technology*. Yuexian Hou, Jian-Yun Nie, Le Sun, Bo Wang and Peng Zhang, Springer Berlin Heidelberg. **7675:** 326-335.

Ramanathan Guha, Vineet Gupta, Vivek Raghunathan and Ramakrishnan Srikant (2015). User Modeling for a Personal Assistant. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, Shanghai, China, ACM. p. 275-284.

Manish Gupta, Rui Li, Zhijun Yin and Jiawei Han (2011). An Overview of Social Tagging and Applications. *Social Network Data Analytics*. Charu C. Aggarwal, Springer US**:** 447-497.

Ronald H Heck and Scott L Thomas (2015). *An Introduction to Multilevel Modeling Techniques: MLM and SEM Approaches Using Mplus*, Routledge.

Mohsen Jamali and Martin Ester (2010). A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the fourth ACM conference on Recommender systems*, Barcelona, Spain, ACM. p. 135-142.

Kalervo Järvelin and Jaana Kekäläinen (2000). IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, Athens, Greece, ACM. p. 41-48.

Victor Lavrenko and W. Bruce Croft (2001). Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New Orleans, Louisiana, USA, ACM. p. 120-127.

Jovian Lin, Kazunari Sugiyama, Min-Yen Kan and Tat-Seng Chua (2013). Addressing cold-start in app recommendation: latent user models constructed from twitter followers. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, Dublin, Ireland, ACM. p. 283-292.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze (2008). *Introduction to Information Retrieval*, Cambridge University Press.

MichaelG Noll and Christoph Meinel (2007). Web Search Personalization Via Social Bookmarking and Tagging. *The Semantic Web*. Karl Aberer, Key-Sun Choi, Natasha Noyet al, Springer Berlin Heidelberg. **4825:** 367-380.

Kazunari Sugiyama, Kenji Hatano and Masatoshi Yoshikawa (2004). Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th international conference on World Wide Web*, New York, NY, USA, ACM. p. 675-684.

Jaime Teevan, Susan T. Dumais and Eric Horvitz (2005). Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, Salvador, Brazil, ACM. p. 449-456.

David Vallet, Iván Cantador and Joemon M Jose (2010). Personalizing web search with folksonomy-based user and document profiles. In *Proceedings of the 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK,* Springer. p. 420-431.

E. M. Voorhees (1999). The TREC-8 Question Answering Track Report. In *Proceedings of the 8th Text REtrieval Conference (TREC-8)*.

Qihua Wang and Hongxia Jin (2010). Exploring online social activities for adaptive search personalization. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, Toronto, ON, Canada, ACM. p.999-1008.

Shengliang Xu, Shenghua Bao, Ben Fei, Zhong Su and Yong Yu (2008). Exploring folksonomy for personalized search. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, Singapore, Singapore, ACM. p. 155-162.

Chengxiang Zhai and John Lafferty (2001). Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, ACM. p. 403-410.

D. Zhou, S. Lawless, J. Liu, S. Zhang and Y. Xu (2015). Query expansion for personalized cross-language information retrieval. In *Proceedings of the 10th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP),* Trento, Italy, IEEE. p. 1-5.

Dong Zhou, Séamus Lawless and Vincent Wade (2012a). Improving search via personalized query expansion using social media. *Information Retrieval,* **15**(3-4): 218-242.

Dong Zhou, Séamus Lawless and Vincent Wade (2012b). Web Search Personalization Using Social Data. In *Proceedings of the Second International Conference on Theory and Practice of Digital Libraries, TPDL 2012, Paphos, Cyprus,* Springer. p. 298-310.

Dong Zhou, Séamus Lawless, Xuan Wu, Wenyu Zhao and Jianxun Liu (2016). A study of user profile representation for personalized cross-language information retrieval. *Aslib Journal of Information Management,* **68**(4): 448-477.

Dong Zhou, Mark Truran, Jianxun Liu, Wei Li and Gareth Jones (2014). Iterative Refinement Methods for Enhanced Information Retrieval. *International Journal of Intelligent Systems,* **29**(4): 341-364.

Arkaitz Zubiaga, Victor Fresno, Ricardo Martinez and Alberto Perez Garcia-Plaza (2013). Harnessing folksonomies to produce a social classification of resources. *IEEE Transactions on Knowledge and Data Engineering,* **25**(8): 1801-1813.

Arkaitz Zubiaga, Alberto Pérez Garcia-Plaza, Víctor Fresno and Ricardo Martinez (2009). Content-based clustering for tag cloud visualization. In *Proceedings of the International Conference on Advances in Social Network Analysis and Mining, ASONAM*, IEEE. p. 316-319.