

# Improving Search via Personalized Query Expansion using Social Media

Dong Zhou, Séamus Lawless, Vincent Wade

**Abstract.** Social tagging systems have gained increasing popularity as a method of annotating and categorizing a wide range of different web resources. Web search that utilizes social tagging data suffers from an extreme example of the vocabulary mismatch problem encountered in traditional Information Retrieval (IR). This is due to the personalized, unrestricted vocabulary that users choose to describe and tag each resource. Previous research has proposed the utilization of query expansion to deal with search in this rather complicated space. However, non-personalized approaches based on relevance feedback and personalized approaches based on co-occurrence statistics only showed limited improvements. This paper proposes a novel query expansion framework based on individual user profiles mined from the annotations and resources the user has marked. The underlying theory is to regularize the smoothness of word associations over a connected graph using a regularizer function on terms extracted from top-ranked documents. The intuition behind the model is the prior assumption of term consistency: *the most appropriate expansion terms for a query are likely to be associated with, and influenced by terms extracted from the documents ranked highly for the initial query*. The framework also simultaneously incorporates annotations and web documents through a Tag-Topic model in a latent graph. The experimental results suggest that the proposed personalized query expansion method can produce better results than both the classical non-personalized search approach and other personalized query expansion methods. Hence, the proposed approach significantly benefits personalized web search by leveraging users' social media data.

**Keywords:** Personalization, Query Expansion, Web Search, Social Media, Tag-Topic Model, Graph Algorithm

---

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation ([www.cngl.ie](http://www.cngl.ie)) at Trinity College Dublin

Dong Zhou, Séamus Lawless, Vincent Wade

*Center for Next Generation Localisation, Knowledge and Data Engineering Group, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland*

{Dong.Zhou, Seamus.Lawless, Vincent.Wade}@scss.tcd.ie

# 1. Introduction

Over the past decade, the area of personalized search has gained much attention in the literature (Cao et al. 2009; Joachims et al. 2005; Chirita et al. 2007; Liu et al. 2004). Providing a personalized service to web search users significantly helps them in satisfying their everyday information needs. Personalized search systems do not retrieve documents that are just relevant to the query but ones that are also relevant to the user's interests; thus different users may actually receive different results for the same query. A key feature of personalized search system is keeping track of the information needs of their users in order to personalize their searches. Therefore, such systems should have a mechanism to learn about their users' search interests. The recorded search interests can then be used to tailor the users' future searches according to their inferred needs.

An important concern in personalized search systems is how to store and represent the gathered usage information. Some systems store this information in an individualized user profile (or user model) (Zhang and Koren 2007; Speretta and Gauch 2005), while other systems maintain an aggregate view of usage information (Agichtein et al. 2006; Smyth and Balfe 2006). Several techniques and data structures can be used to represent user and usage information. They can be broadly classified into two groups: either a vector-based model or a semantic network-based model. A vector-based model (Shen et al. 2005) is made up of a feature vector (or composed of more than one vectors), which is a vector of terms and associated weights. User profiles can also be represented using a semantic network structure (Chirita et al. 2007). In this case the profile is made up of nodes and associated nodes that capture terms and their semantically-related/co-occurring terms respectively. Weights can be assigned to the nodes, their associated nodes, and the links between them. The advantage of this model over a vector-based model is that it can model the relationship between a term and its associated terms.

Personalization in search systems can be achieved by query adaptation, result adaptation, or both. In the result adaptation scenario, search result lists are often re-ranked and/or filtered by incorporating users' interests accordingly (Q. Wang and Jin 2010; Xu et al. 2008). On the other hand, query adaptation attempts to expand (augment) the terms of the user's query with other terms, with the aim of retrieving more relevant results (Chirita et al. 2007; Bertier et al. 2009). Several techniques are used for obtaining terms for query expansion, including relevance feedback-based techniques (Biancalana and Micarelli 2009), co-occurrence-based techniques (Bertier et al. 2009; Chirita et al. 2007), thesaurus-based techniques (Voorhees 1994) among others. In terms of personalized query expansion, additional terms often come from individual user profiles to assist the user in formulating a better query.

Recent years have also witnessed the explosive growth of information on the World Wide Web (WWW). Social media systems have proved a powerful tool to encourage end-user participation in the WWW, for the purpose of categorizing and distributing content, sharing opinions and maintaining relationships. In social tagging systems such as [del.icio.us](http://www.delicious.com)<sup>1</sup>, [LibraryThing](http://www.librarything.com/)<sup>2</sup>, [Flickr](http://www.flickr.com/)<sup>3</sup>, etc., users are able to

---

<sup>1</sup> <http://www.delicious.com>

<sup>2</sup> <http://www.librarything.com/>

<sup>3</sup> <http://www.flickr.com/>

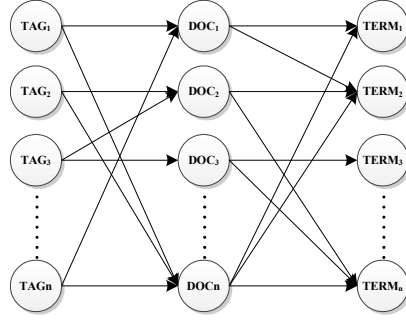
annotate each web resource with any number of free-form tags of their own choice. Such systems have become extremely popular over the past few years. This type of system also provides an ideal test bed for personalized search. A user profile can be easily derived from their feedback, providing a good indication of the user's interests. Moreover, this kind of information is solely based on the user's explicit, public social activities, which means the profile can be safely utilized without disrespecting or compromising the user's privacy.

Social tags produced by users are usually regarded as high quality descriptors of the web pages' topics and a good indicator of web users' interests. Despite this fact, this uncontrolled manner of tagging results in the use of an unrestricted vocabulary. This places a structural barrier between the users and the swathes of globally available information, making searching through the collection difficult and generally less accurate. In current social media systems, search algorithms also tend to be rather simplistic in nature, relying on term matching methods, which often fail to deal with the vocabulary mismatch problem and result in poor ranking results.

To overcome this problem, researchers have attempted to use result re-ranking approaches (Xu et al. 2008; Carmel et al. 2009; Q. Wang and Jin 2010). However, if relevant items could not be fetched in the first place, regardless of the complex re-ranking process, the results still tend to be unsatisfactory.

As previously mentioned, query expansion can partially solve the above mentioned problem. A classic technique is Pseudo relevance feedback (PRF) or local analysis (Baeza-Yates and Ribeiro-Neto 1999), in which expansion terms are automatically extracted from the top-ranked documents and added to the source query, which is then optionally re-weighted (Rocchio 1971). This approach has been previously proven to work well. However, in the context of personalized search, the selected terms may be different from the users' true interests, so that the retrieved documents may not be relevant to a particular user. There have been few attempts at selecting the appropriate expansion terms from a user profile (Biancalana and Micarelli 2009; Bender et al. 2008; Bertier et al. 2009; Chirita et al. 2007). This profile is normally mined from the annotations and content that the user has produced. Past research appears to favor tag-tag relationships, by selecting the most related tags from a user's profile to enhance the source query. Given the fact that the tags might not be the precise descriptions of resources, the resulting retrieval performance has been markedly low (Bender et al. 2008; Bertier et al. 2009). Borrowing from the traditional Information Retrieval (IR) field, local analysis and co-occurrence based user profile representation have been adopted to expand the query according to a user's interaction with the system (Biancalana and Micarelli 2009; Chirita et al. 2007). However, in this case the selection of expansion terms is solely based on lexical matching between the query and the terms which exist in the user profile. If the terms are not found in the user's profile, the query cannot be expanded at all. Furthermore, in local analysis with co-occurrence representation, terms are considered of equal importance when added to the user profile. This may lead to improvements with a global effect rather than on a personalized level.

This paper is concerned with two areas of information retrieval: Search Personalization via individual user profiles and automatic query expansion via a new expansion framework which leverages content from social media applications. Unlike previous approaches to personalized query expansion, which are solely based on lexical matching between the query terms and the terms which exist in the user profile, the method proposed in this paper can expand the query



**Fig. 1. Multiple bipartite graphs between tags, documents and terms**

even if the query terms are not matched with terms in the user's profile. This is achieved by incorporating pseudo-relevance feedback information obtained from top-ranked documents. It is of particular importance to web search which is tailored using social media data because term use tends to be very inconsistent between different users especially when compared with terms appearing in web documents (Golder and Huberman 2005).

The query expansion framework proposed in this paper is based on individual user profiles. In the user profile, terms are modeled according to their relationships, which can be defined by co-occurrence statistics (used as a baseline) or defined by a tag-topic model introduced below. Each term in the user profile will have an associated weighting score calculated based on its relationship with other terms in the profile and terms extracted from top-ranked documents. After calculation, terms with highest scores will be chosen to expand the original query. The underlying theory is to regularize the smoothness of word associations over a connected graph using a regularizer function on terms extracted from top-ranked documents. The intuition behind the model is the prior assumption of term consistency: *the most appropriate expansion terms for a query are likely to be associated with, and influenced by terms extracted from the documents ranked highly for the initial query*. In other words, the selection of expansion terms for a given query is not solely based on lexical matching, but by context enhancing and weighting propagation. If the neighbors of a term in a connected graph are good expansion candidates for a query, then this term is also highly likely to be a good candidate for the query. In addition, the refined weighting scores should be at least somewhat relevant to the enhanced context (i.e. the top-ranked documents retrieved by the initial query, which are assumed to be relevant), which, in our framework, are constrained by a regularizer on the top-ranked documents.

In summary, the motivation to develop this expansion framework is twofold. Firstly, to incorporate pseudo-relevance feedback information obtained from top-ranked documents in the word association graph to expand the initial query. Secondly, traditional personalized query expansion can only work when query terms are found within the user profile, but the framework proposed here will work even when direct lexical matching fails. This fact is particularly important when enhancing search using social media data.

Due to the nature of social tagging systems, tags, web documents and terms extracted from documents are associated in a complex fashion. The relationship between tags and documents can be represented by a tag-document bipartite graph, and the content of documents can be represented by a document-term bipartite graph. Figure 1 shows a sample of these multiple bipartite graphs. In this paper, we propose to simultaneously incorporate the user's annotations and the content information through a statistical model in a latent space graph. This is

achieved by adopting the Author-Topic model introduced by Steyvers et al. (Steyvers et al. 2004), and proposing a Tag-Topic model to learn topic-term and tag-topic distributions from the annotation data in an unsupervised manner. A latent graph is then built based upon the features derived between important terms and tags. The advantage of this approach is that it simultaneously incorporates web page content and annotations, which provides rich information for better performance.

To illustrate the effectiveness of the proposed methodology, we follow previous work on using social data for evaluating personalized search (Xu et al. 2008; Carmel et al. 2009; Q. Wang and Jin 2010) by using data crawled from a large social tagging system. Over two hundred users, distributed worldwide, who are active on the system have been tested. The experimental results suggest that the proposed personalized query expansion method can produce better results than the classical non-personalized search approach and other query expansion methods. The results also demonstrate the effectiveness of modeling content and annotations in the latent graph.

The main contribution of this paper is to propose a regularization framework for query expansion, which aims to produce more accurate personalized retrieval results. The key to expanding the query is the global term consistency over the word graph, which leverages the top-ranked documents retrieved by the query. Another contribution is that the framework also simultaneously incorporates the annotations and web documents through a Tag-Topic model in a latent graph.

The rest of this paper is organized as follows. Related work on query expansion, personalized search, topic models and semi-supervised learning is briefly summarized in Section 2. Section 3 describes the regularization framework for query expansion. Section 4 presents details of how to build the user profiles through a Tag-Topic model. In Section 5 and 6 a report is provided on a series of experiments performed over data crawled from a large social tagging system. This report includes details of the results obtained. Finally, Section 7 concludes the paper and proposes some future work.

## **2. Related Work**

### **2.1 Query Expansion**

Manual query expansion has been studied in early IR systems (Harter 1986). This approach demands user intervention and requires the user to be familiar with the search system, which is generally not true for the modern web. For these reasons, the overwhelming majority of search systems in existence today, function via automatic query expansion. A number of different automatic query expansion techniques have been described in the IR literature. One common technique employs a machine readable thesaurus to locate expansion terms in lists of synonyms (Voorhees 1994). Other approaches extract expansion terms from large collections of documents (Qiu and Frei 1993). Local analysis involving relevance feedback is another popular category of approach. Explicit feedback is often difficult to obtain because users are usually reluctant to provide such information. An alternative method is implicit relevance feedback through pseudo relevance feedback (PRF) (Baeza-Yates and Ribeiro-Neto 1999), in which expansion terms are automatically extracted from the top-ranked documents and added to the source query, which is then optionally re-weighted (Rocchio 1971). This approach has been previously proven to work well. However, in the context of personalized search, the selected terms may be different from the users' true interests, so that

the retrieved documents may not be relevant to a particular user. In contrast, our system ensures that only feedback terms that are relevant to the user's needs are used for query expansion. Web query logs are also used by researchers to bridge the gap between the user-centric query space and author-centric web page space (Cui et al. 2002, 2003). However, in practice, acquiring web query logs is difficult for most researchers due to the various concerns of search companies. Co-occurrence based techniques are also highly attractive, and function by analyzing entire documents (Qiu and Frei 1993), lexical affinity relationships (Carmel et al. 2002) etc. In social tagging systems, these relationships tend to be more complex than web search and hence require more sophisticated modeling methods.

## **2.2 Personalized Search**

Personalized web search has been extensively studied. There are approaches that utilize query log and click-through analysis (Cao et al. 2009; Joachims et al. 2005). There are systems that explore desktop data and external resources (Chirita et al. 2007; Liu et al. 2004). There are also techniques that focus on the user task and activity context (Dou et al. 2007; Luxenburger et al. 2008). This technique is also applied to commercial search engines (Haveliwala et al. 2003).

None of the above work exploits information from social media systems to perform personalized search. In personalized search in social media, the search process is performed over "social" data gathered from Web 2.0 applications such as social bookmarking systems, wikis, blogs, forums and social network sites. Personalization usually involves two general approaches. The first approach runs the unmodified original query for all users but re-ranks the returned results based on an individual user profile. In (Xu et al. 2008) the authors developed a personalization approach to learn about users' interests from their bookmarks and tags, then re-rank the results according to the topic relevance of documents and users' interests. Carmel et al. (Carmel et al. 2009) investigated personalized social search based on the user's social relations. Search results are re-ranked according to their relation to individuals in the user's social network. Wang and Jin (Q. Wang and Jin 2010) explored gathering data from multiple online social systems for adaptive search personalization. They created an interest profile for each user by integrating different streams of social information and then re-ranked the results through a linear combination of different computed scores. 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.

Another group of work modifies or augments a user's original query, this approach is termed query expansion. Researchers have frequently used co-occurring tags to enhance the source query (Bender et al. 2008; Bertier et al. 2009). However, given the fact that the tags might not be precise descriptions of the content, retrieval performance is notably low. Borrowing from the traditional IR field, local analysis has been adopted to expand the query according to a user's interaction with the system (Biancalana and Micarelli 2009; Chirita et al. 2007). In Biancalana and Micarelli's system, the authors used a three-dimensional correlation matrix to build the user profile. Each term of the matrix is linked to an intermediate level extracted from an external resource. The related terms are extracted according to their relevance to the query. However, the selection of expansion terms is solely based on lexical matching between the query and the terms existing in the user profile. If a term is not found, the query cannot be expanded at all. Furthermore, in local analysis, terms are considered of equal

importance when added to the user profile. This may lead to improvements with a global effect rather than on a personalized level. In contrast, the user profile used in the framework proposed by this paper only selects important terms extracted from the annotated content to avoid this problem. The system developed by Biancalana and Micarelli also relies on external categorization, which creates an extra burden and adds uncertainty to the process.

### **2.3 Topic Models**

To learn a latent space graph for use in the framework, this work is also related to the family of topic models. Latent Dirichlet Allocation (LDA), after it was first introduced in (Blei et al. 2003), has quickly become one of the most popular probabilistic text modeling techniques and has inspired much research. There are many extensions to this model, most notably the Author-Topic model (Steyvers et al. 2004), which extracts information about authors and topics from text collections. This model has been adopted in the approach described in this paper and a Tag-Topic model proposed in building user profiles, from which a latent graph can be formed. Some researchers have also employed a modified version of LDA in social networks. Zhou et al. (Ding Zhou et al. 2008) proposed a computationally tractable hierarchical Bayesian network method for modeling social annotations, together with language modeling for personalized ranking. In (Harvey et al. 2011) the authors proposed several hidden topic models to provide more accurate resource ranking. Unlike the modeling processes in their papers, which were deployed on the whole corpus, the model described in this paper is performed on an individual level to ensure a more concrete user profile for query expansion.

### **2.4 Semi-supervised Learning**

The framework described in this paper is influenced by existing work on machine learning, especially graph-based semi-supervised learning (D. Zhou et al. 2004; Zhu et al. 2003; F. Wang et al. 2008). The regularization framework proposed is closely related to label smoothness over the graph. However, the work here is different as the tasks are performed in a different setting. Their tasks are mainly used at the document level such as classification and clustering, while the methods proposed in this paper are focused on the word level and in a query-document dependent setting.

## **3. Personalized Query Expansion Framework**

In this section, the problem addressed by this paper is defined. The personalized query expansion framework is also described in more detail. This method builds upon individual user profiles in which terms are mined from both the annotations a user has made and the resources the user has marked. Each term in the user profile will have an associated weighting score calculated based on its relationship with other terms in the profile and terms extracted from the top-ranked documents. After calculation, the terms with the highest scores will be chosen to expand the original query. The underlying theory is to regularize the smoothness of word associations over a connected graph using a regularizer function on terms extracted from top-ranked documents. In Table 1, we list all the symbols we will use in the algorithms.

**Table 1. Notation used in the personalized query expansion framework and the Tag-Topic Model**

Symbol	Meaning	Symbol	Meaning
<b>Personalized query expansion framework</b>			
$\mathcal{P}$	Social Tagging Data	$\mathcal{D}^{top}$	A set of initial top-ranked documents
$\mathcal{U}$	A finite set of users	$f(w, d)$	Weighting of a word $w$ in a document $d$
$\mathcal{D}$	A finite set of web documents	$V$	A set of nodes in a graph
$\mathcal{T}$	A finite set of tags	$E$	A set of edges in a graph
$\mathcal{A}$	Annotations	$A$	The adjacency matrix of a graph $G$ , $a_{ij}$ is an entry
$t$	A tag	$M$	A diagonal matrix with entries $M_{ii} = \sum_j a_{ij}$
$d$	A document	$S$	A normalized matrix
$q$	A query	$F$	A $n \times c$ matrix where rows represent words and columns represent the documents
$w$	A word	$\mu$	A regularization parameter with $\mu_1 = \frac{1}{1+\mu}$ and $\mu_2 = \frac{\mu}{1+\mu}$
$G$	A connected graph	$I$	An identify matrix
<b>Tag-Topic Model</b>			
$\theta$	Topic-word distributions	$\delta$	Top words for each topic
$\varphi$	Tag-topic distributions	$\mathbf{z}$	Topic Assignments
$\alpha, \beta$	Dirichlet priors	$\mathbf{x}$	Tag Assignments
$\gamma$	The number of expansion words	$L$	The size of lexicon
$\mathcal{O}$	A finite set of topics		

### 3.1 Problem Definition

In social tagging systems such as del.icio.us, users can label interesting web pages with primarily short and unstructured *annotations* in natural language called *tags*. These web pages are denoted as a link to a URL in the del.icio.us website. Textual content is crawled by following a URL link that refers to a *document* or *web document*. Multimedia content is excluded in this research. In response to a *query*, an initial set of the most relevant documents is fetched. We assume that the top “ $c$ ” ranked documents are relevant, and therefore refers to *top-ranked documents*. *Term* refers to a *word* in the vocabulary, these two terminologies are used interchangeably. Terms extracted from documents are specifically called *docTerm*, to be distinguished from general “terms” used in user profiles and from tags.

Formally, social tagging data can be represented by a tuple  $\mathcal{P} := (\mathcal{U}, \mathcal{D}, \mathcal{T}, \mathcal{A})$ , where  $\mathcal{U}, \mathcal{D}, \mathcal{T}$  are finite sets of users, web documents and tags, and  $\mathcal{A} \subseteq \mathcal{U} \times \mathcal{D} \times \mathcal{T}$  is a ternary relation, whose elements are called tag assignments or annotations. The set of annotations of a user is defined as:  $\mathcal{A}_u := \{(t, d) | u, d, t \in \mathcal{A}\}$ . The tag vocabulary of a user, is given as  $\mathcal{T}_u := \{t | (t, d) \in \mathcal{A}_u\}$ . The user’s



set of documents is  $\mathcal{D}_u := \{d | (t, d) \in \mathcal{A}_u\}$ . We further define the docTerm vocabulary of a user to be  $docTerm_u := \{w | w \in \mathcal{D}_u\}$  where  $w$  denotes the words in the document corpus.  $\mathcal{T}_u$  is the full list of tags that the user has used, and  $docTerm_u$  is the vocabulary extracted from the documents that the user has tagged. So that terms in a user profile could be chosen from  $\mathcal{T}_u$ ,  $docTerm_u$ , or  $\mathcal{T}_u \cup docTerm_u$ .

Given a source query  $q$ , a set of terms/words in the user profile  $\{w_1, w_2 \dots w_n\}$ , and a set of initial top-ranked documents  $\mathcal{D}^{top} = \{d_1, d_2 \dots d_c\}$  the goal is to return a ranked list of profile terms to be added to the query, regularized by terms extracted from the top-ranked documents.

Let  $G = (V, E)$  be a connected graph, where nodes  $V$  corresponding to the  $n$  words in the user profile, and edges  $E$  corresponding to the association strengths between words. Further we assume an  $n \times n$  symmetric weight matrix  $A$  on the edges of the graph is given, which  $a_{ij}$  denotes the weight between words  $w_i$  and  $w_j$  and  $M$  is a diagonal matrix with entries  $M_{ii} = \sum_j a_{ij}$ . We also define a  $n \times c$  matrix  $F$  with  $F_{ij} = f(w, d)$  if a word  $w$  is presented in a document  $d$  and  $F_{ij} = 0$  otherwise, where  $f(w, d)$  denotes weighting of  $w$  in  $d$ .

### 3.2 Expansion Framework

Inspired by the semi-supervised learning methods, here we develop a personalized query expansion framework. Formally, the cost function  $\mathfrak{R}(F, w, G)$  in a joint regularization framework similar to (D. Zhou et al. 2004) is defined as:

$$\begin{aligned} \mathfrak{R}(F, w, G) = & \frac{1}{2} \left( \sum_{i,j=1}^n a_{ij} \left\| \frac{f(w, d_i)}{\sqrt{M_{ii}}} - \frac{f(w, d_j)}{\sqrt{M_{jj}}} \right\|^2 \right. \\ & \left. + \mu \sum_{i=1}^n \|f(w, d_i) - f^0(w, d_i)\|^2 \right) \end{aligned}$$

where  $\mu > 0$  is the regularization parameter.  $f^0(w, d_i)$  is the initial weighting matrix of word  $w$  in the document  $d_i$  (*tf-idf* weighting used in the current paper (Jones 1988)). Let  $F$  and  $F^0$  be a refined weighting matrix and an initial weighting matrix, respectively.

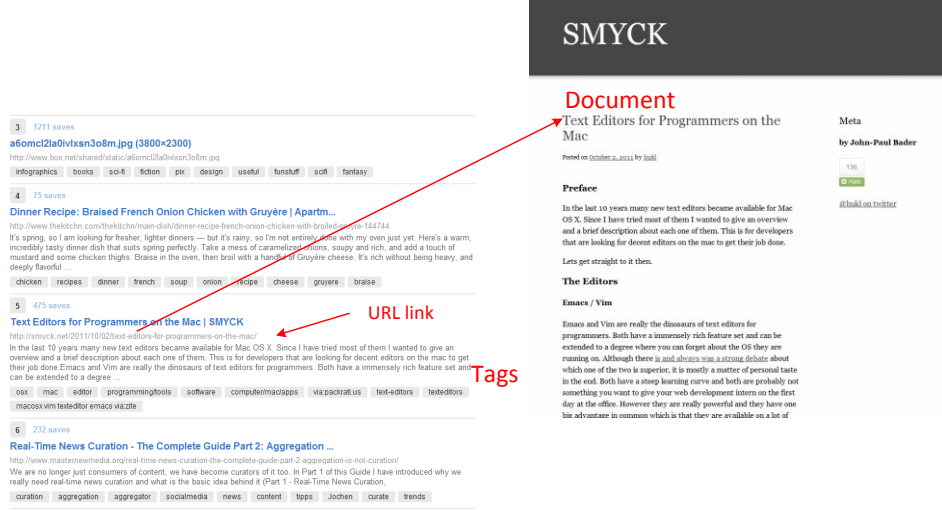
The first term of the right-hand side in the cost function is the global consistency constraint, which means that a good weighting function should not change too much between nearby points. In this paper, nearby points are refined weighting scores with respect to initial relationships between words and context information (top-ranked documents) obtained by initial query. They are likely to have the same effect over the graph. The second term is the fitting constraint, which means the weighting of words should fit the weighting scores of words extracted from the top-ranked documents retrieved by the given query. The trade-off between each other is controlled by the parameter  $\mu$ .

Then the final weighting function is

$$F^* = \arg \min_{F \in \mathbb{R}^{+n}} \mathfrak{R}(F, w, G)$$

After simplifying, a closed form solution can be derived as (see also (D. Zhou et al. 2004; Zhu et al. 2003)):

$$F^* = \mu_2(I - \mu_1 S)^{-1} F^0$$



**Fig. 2. An illustrative example of a social tagging page with one correspondent web document**

where  $\mu_1 = \frac{1}{1+\mu}$ ,  $\mu_2 = \frac{\mu}{1+\mu}$ ,  $S = M^{-\frac{1}{2}}AM^{\frac{1}{2}}$  and  $I$  is an identity matrix. Note that  $S$  is a normalized graph Laplacian matrix. We will introduce the calculation of  $A$  and  $S$  in section 4 when we discuss the user profile construction process. Given the refined weighting matrix  $F$ , the final weighting scores for each word  $w$  could be computed as  $w = \sum_{i=1}^c f(w, d_i)$ , from which we can acquire top  $\gamma$  words from the final ranked list of profile words to be added to the query. It is worth noting that  $\mu_2$  could be eliminated as it does not change the ranking.

An important feature of such computation is that weightings calculated here share similarities with the entries obtained in the  $F^*$  in Zhou et al's paper (D. Zhou et al. 2004), where they try to find largest entry to get the corresponding label while we are trying to find large added weights to select terms. The actual values of weights are not so critical as far as they have discriminative power to separate high potential words from low potential words. They are not used in the later computation. Section 6.2 compares this framework with the method of directly matching terms inside the user profile to illustrate the difference. This concludes our discussion of the proposed framework.

## 4. User Profile Construction

In order to capture accurate information for the construction of the user profile, in this paper the tags and web documents are modeled simultaneously. In addition to Figure 1, an example page with sample tags and a document linked to the marked URL is given in Figure 2. The lexical processing of words inside the document will be detailed in the next section.

In the current paper, an Author-Topic model, introduced by Steyvers et al. (Steyvers et al. 2004), was adopted and a Tag-Topic model was proposed in order to learn topic-word and tag-topic distributions from the annotation data in an unsupervised manner. Then a latent graph is built based upon the features derived. This can be achieved through important docTerms, tags or a mixture of both. The advantage of this approach is that it simultaneously incorporates web page content and annotations, which provides rich information for improved performance.

- 
1. For each tag  $t \in \mathcal{T}$ , choose  $\theta_t \sim \text{Dirichlet}(\alpha)$   
 For each topic  $o \in \mathcal{O}$ , choose  $\varphi_o \sim \text{Dirichlet}(\beta)$
  2. For each document  $d \in \mathcal{D}$   
 Given the vector of tags  $\mathbf{t}_d$   
 For each word  $w_i$  indexed by  $i = 1, \dots, N_d$   
     Conditional on  $\mathbf{t}_d$  choose an tag  $x_i \sim \text{Uniform}(\mathbf{t}_d)$   
     Conditional on  $x_i$  choose a topic  $z_i \sim \text{Discrete}(\theta_{x_i})$   
     Conditional on  $z_i$  choose a word  $w_i \sim \text{Discrete}(\varphi_{z_i})$
- 

**Fig. 3. Generative process of the Tag-Topic model in social tagging data**

#### 4.1 Tag-Topic Modeling for Social Tagging Data

The original author-topic model, introduced by Steyvers et al. (Steyvers et al. 2004), reduces the process of generating a document to a simple series of probabilistic steps. The model not only discovers the topics that are expressed in a document, but also which authors are associated with each topic. To run this model on the social tagging data at an individual user level, we can view the tags as authors in the new proposed model. When generating a document, a tag is chosen at random for each individual word in the document. This tag picks a topic from its multinomial distribution over topics, and then samples a word from the multinomial distribution over words associated with that topic. This process is repeated for all words in the document. This process is summarized in Figure 3, and the graphical model corresponding to this process is shown in Figure 4. Where  $\theta$  and  $\varphi$  are topic-word distributions and tag-topic distributions respectively,  $\alpha$  and  $\beta$  are Dirichlet priors. For each word, the topic and tag assignment are sampled from:

$$p(z_i = j, x_i = k | w_i = m, \mathbf{z}_{-i}, \mathbf{x}_{-i}) \propto \frac{C_{mj}^{\mathcal{W}_o} + \beta}{\sum_{m'} C_{m'j}^{\mathcal{W}_o} + L\beta} \frac{C_{kj}^{\mathcal{T}_o} + \alpha}{\sum_{j'} C_{kj'}^{\mathcal{T}_o} + o\alpha}$$

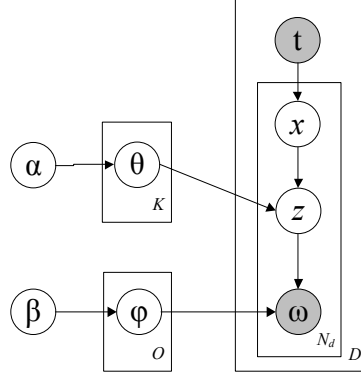
where  $z_i = j$  and  $x_i = k$  represent the assignments of the  $i^{\text{th}}$  word in a document to topic  $j$  and tag  $k$  respectively.  $w_i = m$  represents the observation that the  $i^{\text{th}}$  word is the  $m^{\text{th}}$  word in the lexicon.  $\mathbf{z}_{-i}, \mathbf{x}_{-i}$  represents all topic and tag assignments not including the  $i^{\text{th}}$  word. Furthermore,  $C_{mj}^{\mathcal{W}_o}$  is the number of times word  $m$  is assigned to topic  $j$ , not including the current instance, and  $C_{kj}^{\mathcal{T}_o}$  is the number of times tag  $k$  is assigned to topic  $j$ , not including the current instance,  $L$  is the size of the lexicon.

From the count matrices obtained during the process,  $\theta$  and  $\varphi$  can be easily estimated as:

$$\varphi_{mj} = \frac{C_{mj}^{\mathcal{W}_o} + \beta}{\sum_{m'} C_{m'j}^{\mathcal{W}_o} + L\beta}$$

$$\theta_{kj} = \frac{C_{kj}^{\mathcal{T}_o} + \alpha}{\sum_{j'} C_{kj'}^{\mathcal{T}_o} + o\alpha}$$

where  $\varphi_{mj}$  is the probability of using word  $m$  in topic  $j$ , and  $\theta_{kj}$  is the probability of using topic  $j$  and tag  $k$ . The algorithm assigns words to random



**Fig. 4. Graphical representation for the Tag-Topic model in social tagging data**

topics and tags (from the set of tags annotated to the document), and then repeats the Gibbs sampling process to update topic assignments for several iterations.

The model proposed here is almost identical to the original Author-Topic model except that authors are now replaced by tags. This can be achieved because the mixture weights for different topics are now determined by the tags that a user has assigned to the document, in conjunction with the document corpus. It reflects the user's view of the document, and different documents assigning a same tag will be correctly reflected in the user profile, to ensure more accurate relationships between the terms stored. By learning the parameters of the model, we obtain the set of topics that appear in a corpus and their relevance to different documents, and identify which topics are used by which tag. Hence, the behavior and interests of a user could be identified.

## 4.2 Build the Latent Graph

After the topic-word distributions and tag-topic distributions have been obtained, the adjacency graph of word associations and tag associations can be constructed. Such a task requires the computing of the similarity between tags and/or docTerms. To illustrate how the model could be used in this respect, taking tags as an example, the distance between tags  $t_i$  and  $t_j$  was defined as the symmetric *KL divergence* between the topic distributions conditioned on each of the tags:

$$\text{symKL}(t_i, t_j) = \sum_{o=1}^O \left[ \theta_{io} \log \frac{\theta_{io}}{\theta_{jo}} + \theta_{jo} \log \frac{\theta_{jo}}{\theta_{io}} \right]$$

Similarly we can compute associations between words. The distance between tags and words can also be calculated by using corresponding topic entries in  $\theta$  and  $\phi$ . However, in experiments, this method did not achieve good results compared to using either words or tags alone. This is understandable because the multinomial distribution over words and over tags may contain different meanings. So in the mixture of both approaches, we simply using two latent graphs, by setting their parameters, fitting into the query expansion framework, then combining the words obtained from both.

So a latent graph  $G$  is defined using the latent feature obtained from the tag-topic model, where the nodes denote the terms and the edges  $E$  are weighted by *symKL*. After normalization, matrix  $S = M^{-\frac{1}{2}} A M^{\frac{1}{2}}$  can be calculated. This process is executed offline, and then matrix  $S$  is saved for the query expansion model. Since terms extracted from the documents are not equally important, we

only keep the top  $\delta$  terms from each topic to form the graph. Tags are usually regarded as high quality descriptors of the web pages' topics and a good indicator of web users' interests, so we keep all of them in the user profile. To compare the use of docTerms extracted from documents and tags assigned to the documents for query expansion, three sets of user profiles have been defined: selected docTerms, tags and a mixture of both. In the next section, we will also compare the performance of overall query expansion using this graph and using a co-occurrence based graph. It is also worth noting that according to Wang and Zhang (F. Wang and Zhang 2006), the building of an affinity matrix is very important and can greatly affect the framework effectiveness. So there is no guarantee the latent graph used in this paper is an optimal affinity matrix, but rather a possibility. The evaluation described in the following sections demonstrates that this approach works better than various baselines. Further improvements may be acquired if attempts are made to optimize the matrix employed. This is noted as important future work.

## 5. Evaluation

In the following section a series of experiments are described which have been designed to evaluate the query expansion framework described above. This evaluation focuses on the following thematically related questions:

- (1) Is the proposed personalized query expansion model an improvement over classical non-personalized and query expansion techniques that utilize social media data?
- (2) Will the user profiles built upon the tag-topic model prove an advance over the co-occurrence based model?
- (3) How does the performance differ for users with different amounts of data available on social systems (i.e. active and less active users)?
- (4) Will the filtering of less important words extracted from annotated documents improve the quality of personalized query expansion?
- (5) Are the user profiles containing tags, docTerms and a combination of both equally effective when used in the context of personalized query expansion?

### 5.1 Experimental Data

In order to evaluate these methods on real-world data a crawl was conducted on the popular social tagging site delicious during December 2010. The crawling procedures used by other researchers to download the data from the web (Carman et al. 2008; Harvey et al. 2011) were followed to avoid biased records. To ensure a random sample of recent data the most recent URLs submitted to delicious were downloaded and the usernames of the users who bookmarked them were recorded. After several iterations a sample of 12,043 unique usernames had been collected. Then for each of these usernames the user's bookmarking records were downloaded by analyzing the corresponding web pages. The reason that the delicious API was not used is that restrictions are placed on the volumes of downloads possible, in this case approx. 100 bookmarks. It is the authors' belief that this process can lead to more complete user profiles for users and a more comprehensive test corpus. Non-English users were filtered out based upon their tags because it was desirable to evaluate in a monolingual setting. Also users with less than 10 personal tags were ruled out due to the difficulty in creating a user profile from such few tags. The actual web pages were then crawled, and a total of

**Table 2. Statistics of Delicious dataset sample**

	Delicious
Num of user	5,943
Num of pages	1,190,936
Num of tags	283,339
Avg. num of tags per user	47.68
Avg. num of pages per user	200.39
Max. tags	2,762
Min. tags	10
Max. pages	987
Min. pages	3

5,943 users, 1,190,936 web pages and 283,339 tags were obtained. Table 2 provides statistics which describe the dataset used in the experiments, where, for example, “Max.tags” denotes the maximum number of distinct tags associated with each user.

Four groups of users were created according to the number of bookmarks associated with the users: users with less than 50 bookmarks, denoted as DEL50; users with 50-100 bookmarks, denoted as DEL100; users with 100-500 bookmarks, denoted as DEL500 and finally users with more than 500 bookmarks, denoted as DELgt500. This choice reflects users who are active in the online social system as well as those who are less active, and is consistent with the previous research (Xu et al. 2008; Q. Wang and Jin 2010). 50 randomly selected users from each group together with their tagging records were extracted to form a total collection of 200 test users. The English terms were processed in the usual way, i.e. down-casing the alphabetic characters, removing the stop words and stemming words using the Porter stemmer. All the pre-processed web pages are used in the experiments as the document corpus. No other filtering is conducted. All the information retrieval experiments were performed using the Terrier<sup>4</sup> open source platform.

## 5.2 Evaluation Methodology

For each user, 75% of his/her tags with annotated web pages were used to create the user profile, while the other 25% of his/her tags with annotated web pages were used as a test collection. Tags are subsequently used as queries (see Figure 2 for example tags). Because the query expansion framework described in this paper used individual user profiles to adapt the search, it would be considered personalized search when compared to those methods which rely on textual relevance only (such as *BM25* (Robertson and Zaragoza 2009), used as a non-personalized baseline below).

A subset of users was also randomly selected to train the necessary parameters. Every effort was made to ensure there was no overlap between the training-set of users and the test-set of users.

The major challenge in evaluating a personalized search system is to determine which results are considered relevant and useful to a search query by a specific

<sup>4</sup> <http://www.terrier.org>

- 
- 1: Let  $K$  be the set of words in the user profile that could potentially be added as expansion terms to an input query  $q$ .
  - 2: For each word/term  $w_i$  of  $q$ :
  - 3:  $K \leftarrow K \cup \text{Top}(w)$  where  $\text{Top}(w)$  contains the Top terms with the closest relationship to  $w$  (obtained from co-occurrence matrix by using lexical matching)
  - 4: For each word/term  $w_j$  in  $K$ :
  - 5: Let  $\text{Score}(w_j) \leftarrow \prod_{w_i \in q} (0.01 + \cos(w_i, w_j))$
  - 6: Select top terms of  $K$  with highest scores.
- 

**Fig. 5. Query Expansion procedure based on lexical matching and Co-occurrence based user profiles**

user. We employ the evaluation method used by previous researchers in personalized social search (Xu et al. 2008; Q. Wang and Jin 2010). The main assumption is as follows: Any documents tagged by  $u$  with  $t$  are considered relevant for the personalized query  $(u, t)$  (i.e.  $u$  submits the query  $t$ )<sup>5</sup>. The dependency between personalization and evaluation was also eliminated according to (Carmel et al. 2009)

The following evaluation metrics were chosen to measure the effectiveness of the various approaches: the precision of the top 5 documents ( $P@5$ ), mean reciprocal rank (MRR), mean average precision (MAP) and the recall of the top 5 documents ( $R@5$ ). The first three measurements are commonly used to evaluate search algorithms while the last one is useful for evaluating query expansion systems as this method has been shown to improve both recall and precision in the past. The four metrics were calculated for each user and the mean of all the values was calculated, so that the average performance over test users could be computed. Statistically-significant differences in performance were determined using a paired t-test at a confidence level of 95%.

### 5.3 Experimental Baselines & Runs

In order to usefully evaluate the performance of the personalized query expansion framework 2 different baselines were selected: *BM25* – a popular and quite robust probabilistic retrieval method, and *BM25PRF* – a pseudo-relevance feedback oriented query expansion method based on the Divergence from Randomness (Amati and Rijsbergen 2002) theory. This approach has previously shown good results, which is also a natural choice for evaluating the difference between expanding queries by selecting the terms from the user profiles and from relevant documents<sup>6</sup>.

In addition to non-personalized baselines, we have several personalized baselines. Firstly, a co-occurrence matrix was built according to (Chirita et al. 2007; Biancalana and Micarelli 2009). For all the tags and documents in the training set, we first select the important terms (or keywords) with high *tf-idf* scores (20% used in the experiment). This measurement appears to work better than using document frequency alone (Chirita et al. 2007). We then calculate the

---

<sup>5</sup> See (Xu et al. 2008; Q. Wang and Jin 2010; Carmel et al. 2009) for discussion of pros and cons of this evaluation method.

<sup>6</sup> This method is included in the Terrier distribution.

- 
- 1: For each tag  $t_i$  of  $q$ :
  - 2: For each tag  $t_j$  in the user profiles:
  - 3: Let  $Score(t_j) \leftarrow J(t_i, t_j)$
  - 4: Select top tags in the user profiles with highest scores.
- 

**Fig. 6. Query expansion procedure based on co-tagging**

cosine similarity between two words  $w_i$  and  $w_j$  by using:

$$cos(w_i, w_j) = \frac{DF_{w_i, w_j}}{\sqrt{DF_{w_i} \cdot DF_{w_j}}}$$

Where  $DF_{w_i}$  is the document frequency of word  $w_i$ . Note that here tags and docTerms are modeled into the matrix together. After obtaining this matrix, we use it in the following two ways:

**Lexical matching.** Processing the user profile by using lexical matching between query terms and terms that exist in the user profile. This is achieved by the procedure described in Figure 5 with an additional operation to calculate the correlation between the word and all words/terms in the submitted query. This method will leave some queries un-expanded if no matching has been found in the user profile. This method is denoted as *LMCO* later on.

**Query Expansion Framework.** In a similar fashion to how the latent graph is used in the framework, graph  $G$  is used, where the nodes denote the terms and the edges  $E$  are weighted by their co-occurrence similarity. After normalization, matrix  $S$  is used as usual way for the query expansion model. We denote this method as *PQECO*.

We are also interested in the personalization approach proposed by (Bender et al. 2008) which is based on pure tag-tag relationships. So an additional baseline is included here based on the co-tagging activities a user performed. In this case, the user profiles contain training tags with their co-tagging statistics computing using the Jaccard coefficient:

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

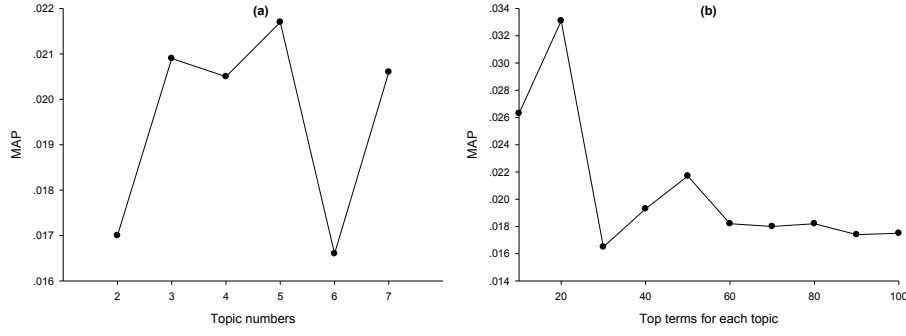
Where  $|N_{t_i}|$  denotes the number of documents tagged by  $t_i$  and  $|N_{t_i} \cap N_{t_j}|$  denotes the number of documents co-tagged by  $t_i$  and  $t_j$ . This method is described in detail in Figure 6 and denoted as *COTAG* hereafter.

Finally, there are also three variants of the proposed query expansion method depending on the characteristics of the user profiles. For those user profiles which only contain terms extracted from the documents, the algorithm is denoted as *PQE\_terms*, for those user profiles which only contain tags, the algorithm is denoted as *PQE\_tags*. *PQE\_mix* is used to represent the query expansion method which uses user profiles that containing a mixture of terms and tags.

## 5.4 Parameter Setting

One important part of the experiment involved establishing the optimal values for the various parameters discussed above. The following section describes in detail how values for  $\phi$  &  $\delta$  (components of the user profile construction) and  $\mu$  &  $\gamma$  (important elements in the query expansion process) were selected. While there are still many parameter settings in addition to these four in the algorithms used





**Fig. 7. The impact of varying parameters  $O$  (a) &  $\delta$  (b)**

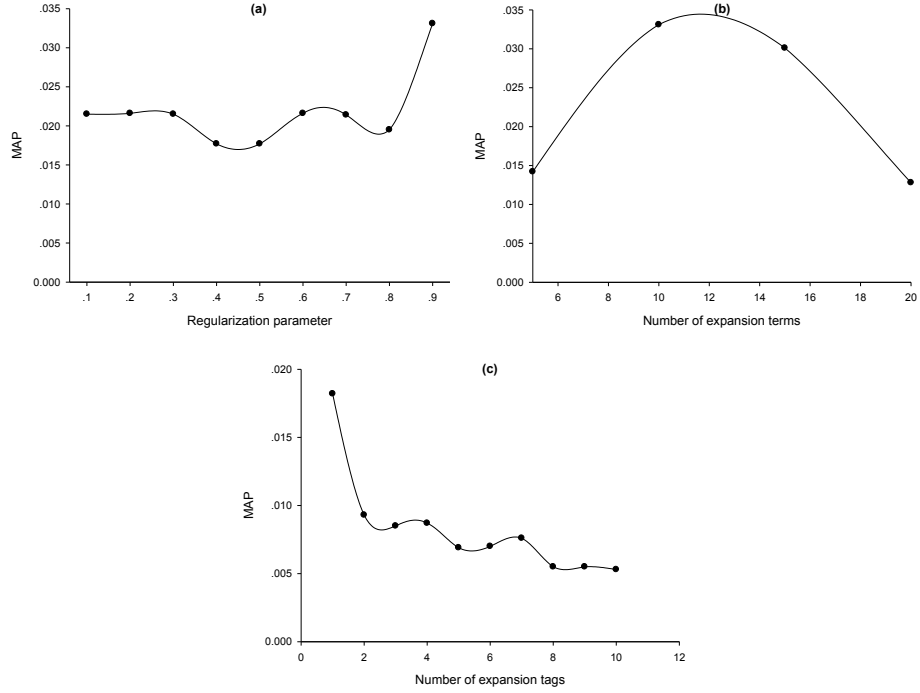
by this research, they are only briefly mentioned in the following section as they are deemed less important.

#### 5.4.1 Setting the tag-topic modeling

The topic numbers  $O$  and the top terms for each topic  $\delta$  are used in the user profile creation process to control the association accuracy and number of terms in the profile. The selection of topic numbers is illustrated by Figure 7(a), the best value of this parameter in terms of MAP was “5” with fewer topics preferred. This partially predicts the power of latent features in the framework. We did not test topic numbers exceeding 10 because of two reasons: firstly, low numbers of topics tend to work better; secondly, although the dataset is large, for each user this parameter tends to be small as it only contains a limited number of tags and document terms. Testing this feature would be highly desirable if the user profiles in the collection were enlarged, which is planned for future work. A number of runs were executed with a spread of settings from 10 to 100 for the parameter  $\delta$ . As shown in Figure 7(b), interestingly the best values were obtained when the number of words per topic is 20. This further confirms the argument above that not all terms should be considered of equal importance when being added to the user profile as was the case in previous research (Biancalana and Micarelli 2009). 20 was selected as the value for  $\delta$  in all subsequent experiments.

#### 5.4.2 Setting the query expansion framework

Previous work on the regularization parameter  $\mu_1$  suggests that a larger value will give optimal performance (D. Zhou et al. 2004). To verify this finding, we conducted a number of runs against the training-set of users, with a spread of settings for the  $\mu_1$  parameter. As illustrated by Figure 8(a), the highest MAP scores were obtained when  $\mu_1 = 0.9$ . In general, the number of expansion terms  $\gamma$  should be within a reasonable range in order to produce consistently good performance. Too many expansion terms not only consumes more time in the retrieval process, but can also have side-effects on the retrieval performance. This is even worse when choosing how many tags should be added to the query. We examine the performance of the query expansion by using 5, 10, 15, and 20 expansion terms and 1 to 10 expansion tags on the training data. The results are shown in Figure 8(b) for terms and 8(c) for tags. The best performance is obtained with around 10 terms and 1 tag. It is worth noting that the curve produced by tags is flatter than that produced by terms. The curve of the tags demonstrated a steady drop after 1 tag. This was attributed to the fact that using tags alone for expanding



**Fig. 8. The impact of varying parameters  $\mu_1$  (a),  $\gamma$  for terms (b) &  $\gamma$  for tags (c)**

queries is not sufficient to acquire optimal performance. It was decided to set  $\mu_1 = 0.9$  and  $\gamma = 10$  for terms and  $\gamma = 1$  for tags in all subsequent experiments.

#### 5.4.3 Setting other parameters

The number of top documents  $\mathcal{D}^{top}$  used in the query expansion framework was set to 10 empirically (from 100 retrieved documents) and the hyper-parameters to run the topic model were set to  $\alpha = 0.01$  and  $\beta = 0.1$ . The parameters for *BM25PRF* were also set to 10 documents and 10 terms as they gave the highest retrieval effectiveness. The *BM25* parameters were left unchanged from their default values in the Terrier distribution. Finally the number of expansion terms applied to *LMCO* and *COTAG* were set to 10 and 5 as these provided the best performance.

## 6. Experimental Results

In this section we present our experimental results. The personalized model is first compared to the non-personalized model. Then we evaluate the effectiveness of the proposed query expansion framework with personalized query expansion based on lexical matching, the effectiveness of using the Tag-topic model is also compared to user profiles built by using co-occurrence statistics. Finally, we illustrate the performance of the proposed model with expansion terms obtained by pure co-tagging activities. We conclude this section by detailing a comparison between different groups of users.

**Table 3. Performance of the six personalized query expansion runs and two non-personalized baselines**

	MAP	MRR	P@5	R@5
<i>BM25</i>	0.0354	0.0411	0.0128	0.0483
<i>BM25PRF</i>	0.0278	0.034	0.0113	0.0398
<i>LMCO</i>	0.0341	0.0432	0.0129	0.0471
<i>COTAG</i>	0.041	0.046	0.014	0.0485
<i>PQECO</i>	0.042	0.0498	0.015	0.0533
<i>PQE_terms</i>	0.0413†*	0.0495†*	0.0155†*	0.0543†*
<i>PQE_tags</i>	0.038*	0.0456†*	0.0138†*	0.0506†*
<i>PQE_mix</i>	0.0448†* <sup>pcl</sup>	0.053†* <sup>pcl</sup>	0.0158†* <sup>pcl</sup>	0.0574†* <sup>pcl</sup>

## 6.1 Personalization vs Non-personalization

### 6.1.1 Overall performance

This set of experimental results describes the performance of the three personalized query expansion runs proposed in this paper together with two non-personalized baselines on the overall test users, which are shown in Table 3. The statistically significant differences are marked as † w.r.t to the *BM25* baseline and \* w.r.t to the *BM25PRF* baseline.

As illustrated by the results, the *BM25PRF* model was the lowest performer for all evaluation metrics. This result is not surprising because the evaluation described in this paper is based upon a personalized-approach rather than the non-personalized evaluation model normally employed in the large evaluation campaigns. This further demonstrates that merely borrowing common techniques from traditional IR will not solve the personalized search problem. Pleasingly, the three personalized query expansion-based search models all outperform the simpler text retrieval model with the highest improvement of 28.95% (In terms of the *PQE\_mix* method with the MRR metric when compared to *BM25*), which is statistically significant. It should be noted that all three personalized query expansion methods provide an average improvement of 40.74% compared to the traditional query expansion method with the highest improvement at 61.15%, which is also statistically significant.

There were noticeable improvements in retrieval effectiveness when using user profiles which consisted of terms and a mixture of terms and tags in query expansion, but a more modest increase for the user profiles which consisted of tags alone. This reinforces the earlier finding that using tags alone for expanding queries is not sufficient. Another exciting observation is that in many cases, the personalized query expansion methods, even though tuned for MAP, can outperform the baselines for all the evaluation metrics, with statistically significant improvements in almost entire runs. We will delay the discussion of recall results (R@5) until later in this section.

Algorithm	<i>PQE terms</i>	<i>BM25PRF</i>	<i>PQE terms</i>	<i>BM25PRF</i>	<i>PQE terms</i>	<i>BM25PRF</i>	<i>PQE terms</i>	<i>BM25PRF</i>
Base Term	tool	tool	logo	logo	swap	swap	food	food
Expansion terms	com	audio	design	busi	game	512	comment	list
	page	collabor	view	compani	com	disk	de	groceri
	media	convert	web	corpor	dai	mb	imag	health
	applic	creativ	page	design	video	memori	book	herb
	free	cut	post	firm	free	partit	photo	market
	blog	exchang	comment	fsc	onlin	ram	number	natur
	design	pro	http	ident	product	sw	video	organ
	share	resourc	mobil	green	list	swap	collect	store
	social	man	www	logotyp	swap	swapon	chicken	supplement
	market	add	intel	vector	make	swappi	photograph	vitamin

Fig. 9. Example expansion terms for the selected queries using *PQE\_terms* and *BM25PRF*

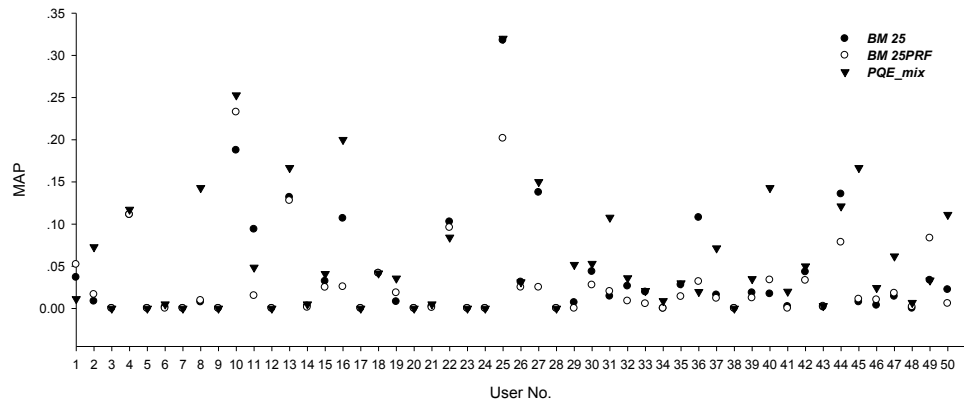


Fig. 10. Individual analysis

### 6.1.2 Expansion terms differences

In addition to the overall performance evaluation, a side-by-side comparison of example expansion terms is also depicted for selected terms by using the *PQE\_terms* and *BM25PRF* models in Figure 9. The two methods produced very different sets of expansion terms for a query. The personalized query expansion method presents many good personalized recommended terms closely resembling the user’s interests while the *BM25PRF* model generates terms more closely aligned with the corpus statistics. For example, the suggested terms for the query “logo” produced by the personalized model are specifically about logo design for particular companies (expansion terms: “design”, “web”, “page”, “mobil”, “intel”<sup>7</sup>, etc., for example) while the terms extracted by the *BM25PRF* method are actually about more general company logo design (expansion terms: “busi”, “company”, “corpor”, “design”, etc.). Similar differences can be found in other cases, for example, with regard to the source query “swap” the *BM25PRF* method selected terms which are about swapping hardware (expansion terms: “disk”, “mb”, “memori”, etc.) while the user is in fact interested in articles talking about swapping online games (expansion terms: “game”, “free”, “onlin”, “product”, etc.).

<sup>7</sup> These terms are stemmed words.

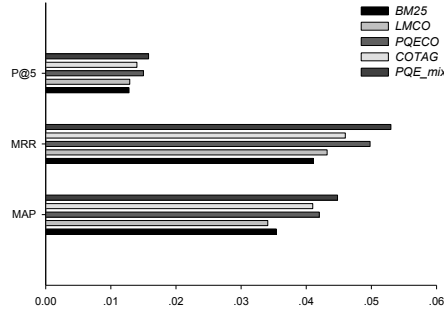


Fig. 11. Comparison with personalized baselines

### 6.1.3 Individual Analysis

To further explain the success of the personalized query expansion methods at an individual user level, a separate figure is presented which depicts the performance of all the users in the DEL50 group. As shown in figure 10, personalized runs outperformed non-personalized runs on many users (around 62%). Only 12% of users have non-personalized baselines which outperformed the personalized approach. This indicates that for most users, it is beneficial to personalize their search results.

## 6.2 Personalized Query Expansion using Lexical Matching and Co-occurrence statistics

The goal of the second set of experiments is to evaluate the performance of personalized query expansion using lexical matching and co-occurrence statistics (*LMCO*), in comparison with the proposed framework (*PQECO*). Experimental results are shown in Table 3 and visualized in Figure 11, together with the non-personalized baseline *BM25* and the results obtained by using the *PQE\_mix* method, which tends to acquire the best performance in the last subsection. The statistically significant differences in the table are marked as *l* w.r.t to the *LMCO* baseline and *p* w.r.t to the *PQECO* for the *PQE\_mix* method only.

As we can see from the figure, query expansion solely based on co-occurrence statistics and lexical matching is unsatisfactory. Although the performance is better in terms of MRR and P@5 metrics when compared to the non-personalized baseline, however, in the MAP metric the performance is even lower. After examining the expanded terms in the *LMCO* model, it was found that because of the nature of social tagging systems, many tags are freely chosen and different from the terms stored in the user profiles, leave a large number of queries un-expanded. Furthermore, the expanded terms sometimes show noise, resulting in lower performance than the *BM25* baseline.

However, using the same co-occurrence matrix as in *LMCO*, the *PQECO* method works much better, with performance just slightly lower than the *PQE\_mix* method. In some metrics it appears to work better than *PQE\_terms*. This shows the power of using pseudo-relevance feedback documents to enhance the word graphs. Also the effectiveness of using the Tag-Topic model is also empirically confirmed (in terms of *PQE\_mix* which works better than *PQECO*). It should be noted that the improvements achieved by *PQE\_mix* in comparison to *PQECO* and *LMCO* are statistically significant.

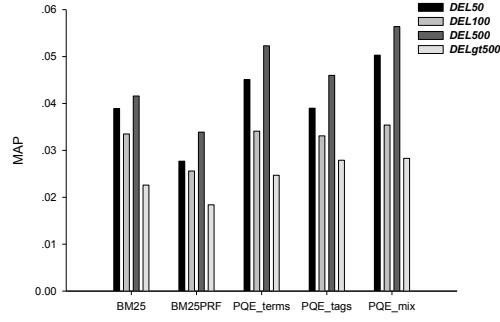


Fig. 12. Comparison between different user groups

### 6.3 Comparison with co-tagging

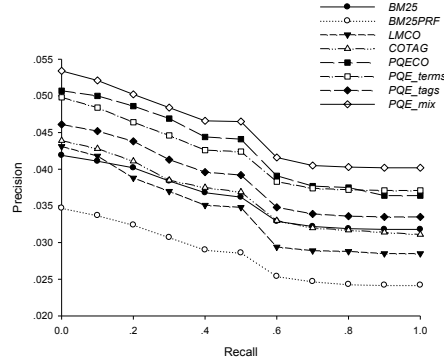
We now examine the performance of the proposed model with expansion terms obtained by pure co-tagging activities. This is also demonstrated in Table 3 (statistically significant differences are marked as *c* for *PQE\_mix* only) and in Figure 11. Surprisingly, this approach outperforms the method that uses lexical matching and co-occurrence statistics (*LMCO*). As these two methods share some similarities, for example, if terms in the query are not found in the user profile, the query will remain unexpanded in the final run, we can only draw the conclusion that tags have a more sound effect in personalized query expansion exercises. The results also confirm that the proposed personalized query expansion framework outperforms all the personalized query expansion baselines used in the experiments described in this paper, with the highest improvement reaching 31.38% in terms of MAP (*PQE\_mix* against *LMCO*). The use of the Tag-Topic model in the framework leads to a highest improvements of 6.67% in terms of MAP over the use of the co-occurrence matrix in the framework.

### 6.4 Comparison between groups

Next, the performance differences between different groups of users are considered. Figure 12 shows a comparison of performance for the four groups in the context of user activities with the MAP metric (as similar behavior is observed in other metrics). As can be seen, the results are mixed. Still, the personalized query expansion methods outperform the non-personalized approaches except in one case where *PQE\_tags* performed worse than the baseline in the DEL100 group. However, the differences between the *PQE\_tags* and *PQE\_terms* methods among different groups of users vary, for example, *PQE\_tags* achieves better MAP in the DELgt500 group than the *PQE\_terms* method, but the same is not true for other groups. The most effective method is *PQE\_mix*, which is better than the other methods both for users who are more active on social systems and those who are less active. Moreover, the improved performance on low active users does demonstrate that this method is very effective even for users who do not have much available data in online social tagging systems.

### 6.5 Recall Results

In addition to the precision-based measurements, the personalized query expansion methods also showed significant improvements with regard to the recall-based metric  $R@5$ . These improvements are on a similar scale when



**Fig. 13. Recall-precision plots for the eight retrieval systems**

compared to the non-personalized and various personalized baselines. This reveals the benefits of adopting this framework even in situations where recall is important. An illustration of these findings is supplied in Figure 13, which gathers together recall-precision plots for the five retrieval systems. As can be seen in this figure, the improvements of the proposed personalized methods are consistent. Generally speaking, the personalized approaches can achieve better recall performance than non-personalized approaches.

## 7. Conclusion and Future Work

In this article a novel query expansion framework was described which is based on individual user profiles mined from the annotations and resources the users bookmarked. The intuition behind the model is the prior assumption of term consistency: *the most appropriate expansion terms for a query are likely to be associated with, and influenced by terms extracted from the documents ranked highly for the initial query*. A tag-topic model was also introduced which simultaneously integrates the annotations and web documents through a statistical model in a latent space graph. The proposed personalized technique performed well on the social data crawled from the web, delivering statistically significant improvements over non-personalized and personalized representative baseline systems with improvements up to 61.15% and 31.38% respectively. The effectiveness of the approach was also demonstrated for users with different levels of tagging activity and at an individual user level.

This research continues along several dimensions. The proposed framework is a general query expansion framework, not only should it utilize social media data, but also data obtained through web search logs. Future work is currently being planned to apply this framework to a large sized, real-world web search log to test its effectiveness. Also, an optimization algorithm is being designed for incremental updates of the user profile generation and optimal graph generation. Latent Semantic Indexing (LSI) (Deerwester et al. 1990) is an alternative to the topic models used to calculate the semantic information, an evaluation of this approach will be included in future work to compare performance. Future work will also include the integration of query expansion and results re-ranking. Information from the user profile is used to re-rank feedback documents, and then a subset of the documents that are most relevant to the user can be selected and terms can be extracted to be added to the query.

## References

1. Agichtein, E., Brill, E., & Dumais, S. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, USA, 2006* (pp. 19-26). 1148177: ACM. doi:10.1145/1148170.1148177.
2. Amati, G., & Rijsbergen, C. J. V. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4), 357-389, doi:10.1145/582415.582416.
3. Baeza-Yates, R. A., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*: Addison-Wesley Longman Publishing Co., Inc.
4. Bender, M., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Parreira, J. X., et al. Exploiting social relations for query expansion and result ranking. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on, 7-12 April 2008 2008* (pp. 501-506)
5. Bertier, M., Guerraoui, R., Leroy, V., & Kermarrec, A.-M. Toward personalized query expansion. In *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems, Nuremberg, Germany, 2009* (pp. 7-12). 1578004: ACM. doi:10.1145/1578002.1578004.
6. Biancalana, C., & Micarelli, A. 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, 2009* (pp. 1060-1065). 1633656: IEEE Computer Society. doi:10.1109/cse.2009.492.
7. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3, 993-1022, doi:10.1162/jmlr.2003.3.4-5.993.
8. Cao, H., Jiang, D., Pei, J., Chen, E., & Li, H. Towards context-aware search by learning a very large variable length hidden markov model from search logs. In *Proceedings of the 18th international conference on World wide web, Madrid, Spain, 2009* (pp. 191-200). 1526736: ACM. doi:10.1145/1526709.1526736.
9. Carman, M. J., Baillie, M., & Crestani, F. Tag data and personalized information retrieval. In *Proceeding of the 2008 ACM workshop on Search in social media, Napa Valley, California, USA, 2008* (pp. 27-34). 1458591: ACM. doi:10.1145/1458583.1458591.
10. Carmel, D., Farchi, E., Petruschka, Y., & Soffer, A. Automatic query wefinement using lexical affinities with maximal information gain. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland, 2002* (pp. 283-290). 564427: ACM. doi:10.1145/564376.564427.
11. Carmel, D., Zwerdling, N., Guy, I., Ofek-Koifman, S., Har'el, N., Ronen, I., et al. 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, 2009* (pp. 1227-1236). 1646109: ACM. doi:10.1145/1645953.1646109.
12. Chirita, P.-A., Firan, C. S., & Nejdl, W. 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, 2007* (pp. 7-14). 1277746: ACM. doi:10.1145/1277741.1277746.
13. Cui, H., Wen, J.-R., Nie, J.-Y., & Ma, W.-Y. Probabilistic query expansion using query logs. In *Proceedings of the 11th international conference on World Wide Web, Honolulu, Hawaii, USA, 2002* (pp. 325-332). 511489: ACM. doi:10.1145/511446.511489.
14. Cui, H., Wen, J.-R., Nie, J.-Y., & Ma, W.-Y. (2003). Query Expansion by Mining User Logs. *IEEE Trans. on Knowl. and Data Eng.*, 15(4), 829-839, doi:10.1109/tkde.2003.1209002.
15. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6), 391-407.
16. Dou, Z., Song, R., & Wen, J.-R. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada, 2007* (pp. 581-590). 1242651: ACM. doi:10.1145/1242572.1242651.
17. Golder, S., & Huberman, B. A. (2005). The structure of collaborative tagging systems. *Journal of Information Science*, 32, 198-208.
18. Harter, S. P. (1986). *Online information retrieval: concepts, principles, and techniques*: Academic Press Professional, Inc.
19. Harvey, M., Ruthven, I., & Carman, M. J. Improving social bookmark search using personalised latent variable language models. In *Proceedings of the fourth ACM*



- international conference on Web search and data mining, Hong Kong, China, 2011* (pp. 485-494). 1935898: ACM. doi:10.1145/1935826.1935898.
20. Haveliwala, T., Kamvar, S., & Jeh, G. (2003). An Analytical Comparison of Approaches to Personalizing PageRank. Stanford InfoLab.
  21. Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, Salvador, Brazil, 2005* (pp. 154-161). 1076063: ACM. doi:10.1145/1076034.1076063.
  22. Jones, K. S. (1988). A statistical interpretation of term specificity and its application in retrieval. In *Document retrieval systems* (pp. 132-142): Taylor Graham Publishing.
  23. Liu, F., Yu, C., & Meng, W. (2004). Personalized Web Search For Improving Retrieval Effectiveness. *IEEE Trans. on Knowl. and Data Eng.*, 16(1), 28-40, doi:10.1109/tkde.2004.1264820.
  24. Luxemburger, J., Elbassuoni, S., & Weikum, G. Matching task profiles and user needs in personalized web search. In *Proceeding of the 17th ACM conference on Information and knowledge management, Napa Valley, California, USA, 2008* (pp. 689-698). 1458175: ACM. doi:10.1145/1458082.1458175.
  25. Qiu, Y., & Frei, H.-P. Concept based query expansion. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, Pittsburgh, Pennsylvania, United States, 1993* (pp. 160-169). 160713: ACM. doi:10.1145/160688.160713.
  26. Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.*, 3(4), 333-389, doi:10.1561/15000000019.
  27. Rocchio, J. (1971). Relevance Feedback in Information Retrieval. In *The SMART Retrieval System* (pp. 313-323).
  28. Shen, X., Tan, B., & Zhai, C. Implicit user modeling for personalized search. In *Proceedings of the 14th ACM international conference on Information and knowledge management, Bremen, Germany, 2005* (pp. 824-831). 1099747: ACM. doi:10.1145/1099554.1099747.
  29. Smyth, B., & Balfe, E. (2006). Anonymous personalization in collaborative web search. *Inf. Retr.*, 9(2), 165-190, doi:10.1007/s10791-006-7148-z.
  30. Speretta, M., & Gauch, S. Personalized Search Based on User Search Histories. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, 2005* (pp. 622-628). 1092551: IEEE Computer Society. doi:10.1109/wi.2005.114.
  31. Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, WA, USA, 2004* (pp. 306-315). 1014087: ACM. doi:10.1145/1014052.1014087.
  32. Voorhees, E. M. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Dublin, Ireland, 1994* (pp. 61-69). 188508: Springer-Verlag New York, Inc.
  33. Wang, F., Li, T., Wang, G., & Zhang, C. Semi-supervised classification using local and global regularization. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 2, Chicago, Illinois, 2008* (pp. 726-731). 1620185: AAAI Press
  34. Wang, F., & Zhang, C. Label propagation through linear neighborhoods. In *Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, 2006* (pp. 985-992). 1143968: ACM. doi:10.1145/1143844.1143968.
  35. Wang, Q., & Jin, H. Exploring online social activities for adaptive search personalization. In *Proceedings of the 19th ACM international conference on Information and knowledge management, Toronto, ON, Canada, 2010* (pp. 999-1008). 1871564: ACM. doi:10.1145/1871437.1871564.
  36. Xu, S., Bao, S., Fei, B., Su, Z., & Yu, Y. Exploring folksonomy for personalized search. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore, Singapore, 2008* (pp. 155-162). 1390363: ACM. doi:10.1145/1390334.1390363.
  37. Zhang, Y., & Koren, J. Efficient bayesian hierarchical user modeling for recommendation system. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, The Netherlands, 2007* (pp. 47-54). 1277752: ACM. doi:10.1145/1277741.1277752.

38. Zhou, D., Bian, J., Zheng, S., Zha, H., & Giles, C. L. Exploring social annotations for information retrieval. In *Proceeding of the 17th international conference on World Wide Web, Beijing, China, 2008* (pp. 715-724). 1367594: ACM. doi:10.1145/1367497.1367594.
39. Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Scholkopf, B. (2004). Learning with local and global consistency. *Advances in Neural Information Processing Systems, 16*, 321-328, doi:citeulike-article-id:922481.
40. Zhu, X., Ghahramani, Z., & Lafferty, J. Semi-supervised learning using Gaussian fields and harmonic functions. In *In The 20th International Conference on Machine Learning (ICML), 2003* (pp. 912-919). doi:citeulike-article-id:3110274.