Soft Computing (2020) 24:1707–1716
https://doi.org/10.1007/s00500-019-03998-1

METHODOLOGIES AND APPLICATION

Integrating social annotations into topic models for personalized document retrieval

Bo Xu[1] · Hongfei Lin[1] · Yuan Lin[1] · Yizhou Guan[1]

Published online: 24 April 2019

Abstract
Social annotations are valuable resources generated by users on the Web, which encode abundant information on user preferences for certain documents. Social annotation-based information retrieval has been studied in recent years for personalizing search results and fulfilling user information needs. However, since social annotations are complicated and associated with users, documents and tags simultaneously, it remains a great challenge to fully capture the potentially useful information for improving retrieval performance. To meet the challenge, we propose a novel method to integrate social annotations into topic models for personalized document retrieval. Our method first reconstructs candidate documents for a given query using social tags of documents to capture user preferences. The reconstructed documents are tailored to user preferences for achieving better performance. We then generalize the latent Dirichlet allocation-based topic models by considering the relationship among users, social tags and documents from social annotations. The modified topic model optimizes the distribution of latent topics of documents for different users to meet user information needs. Experimental results show that our method can significantly outperform the state-of-the-art baseline models for improving the performance of personalized retrieval.

Keywords Social annotations · Document reconstruction · Topic models · Document retrieval

1 Introduction

The development of the Internet facilitates people's life by establishing online social relationship through Web applications, such as microblogs, Web social networks and social annotations. Online social relationship, as valuable resources of users, attracts much attention of researchers and practitioners for user profiling and personalized services. Related research has been carried out based on microblogs and social networks focusing on various aspects, such as discovering latent communities, detecting recent topics from temporal text streams, and retrieving highly dynamic information.

Social annotations, as one type of social network knowledge resources, play a crucial role in existing research. Social annotations are user generated tags for labeling resources such as Web pages, links and documents, which encode user preferences for certain items. Web applications, such as *delicious* (del.icio.us) and *flickr* (flikckr.com), rely on social annotations to provide personalized services. These applications attract much attention of researchers focusing on the usage patterns of tagging system (Golder and Huberman 2006) and the detection of hidden semantics (Wu et al. 2006). These studies indicate that social annotations capture the complex relationship between users and web pages through various tags, which are useful for personalized recommendation and retrieval. However, social annotations have not been fully investigated due to the complex relationship of users and resources. This information correlates social users with their interested resources through tags and can be highly useful for modeling user preferences in personalized document retrieval (Zhou et al. 2008). We focus on the usage of social annotations for personalizing search results in this study.

Existing studies have shown that social annotations are effective in improving information retrieval performance of different related tasks (Zhou et al. 2008; Hotho et al. 2006; Bouadjenek et al. 2013; Pantel et al. 2012; Lin et al. 2011). In these studies, social annotations are mostly used to generate user-oriented features for capturing user preferences in the

Communicated by V. Loia.

✉ Bo Xu
xubo@dlut.edu.cn

✉ Hongfei Lin
hflin@dlut.edu.cn

[1] Dalian University of Technology, Room A923 of Chuangxinyuan Building, Dalian, China

search process. For example, Zhou et al. (2008) proposed a new generative model using a computationally tractable hierarchical Bayesian network to enhance document and query language models with social annotations. Hotho et al. (2006) presented a formal model and a new search algorithm for folksonomies, called FolkRank, which used the structure of folksonomy to enhance retrieval performance. Bouadjenek et al. (2013) proposed a personalized document representation approach for document ranking based on social activities of users. Pantel et al. (2012) studied the utility of social annotations in relevance-based retrieval, which largely facilitated future research on personalized document retrieval. Lin et al. (2011) adopted social annotations as the source of query expansion terms for learning to rank a set of terms and better understanding the query intents. However, user preferences are still partly ignored in these studies, which may further enhance the retrieval performance for better fulfilling the information needs of different users.

To fully integrate user preferences in document retrieval, we investigate social annotations under the framework of topic models to enhance personalized document retrieval. Topic models have been widely used in information retrieval tasks, which can accurately match user query and candidate documents at the topic level for producing more complete and satisfactory search results. Latent Dirichlet allocation (LDA) is a generative probabilistic topic model and has been widely used to identify latent topics of documents within large corpus. Therefore, we believe that LDA may help to capture useful personalized information from social annotations for document retrieval.

In the paper, we propose a novel method based on social annotations for personalized document retrieval. Our method makes the most of social annotations in topic model-based retrieval from two respects: document reconstruction and topic optimization. Document reconstruction modifies original documents based on user tags to highlight useful words and reduce useless words toward information needs of users. Topic optimization improves document-topic distribution obtained from LDA to consider user interests. Finally, we incorporate the optimized topic models into query likelihood language retrieval model and conduct retrieval based on the reconstructed documents for high personalized retrieval performance. Experimental results show that our method is effective in improving the retrieval performance in comparison with the state-of-the-art baseline methods. We summarize the main contributions of this work as follows.

(1) We propose to use social annotations to reconstruct candidate documents for retrieval, which highlights the key words and reduce the redundant words in measuring the relevance of documents.
(2) We propose to modify the topic distribution produced by LDA with social annotations, which increases the influence of user preferences on topic detection for better matching candidate documents.
(3) We conduct experiments on social annotation-based retrieval dataset, and experimental results show that our method outperforms the baseline methods to a large extent.

The rest of the paper is organized as follows. Section 2 introduces the related research work. Section 3 details the proposed method and provides insights into our method. Section 4 reports the experimental results with in-depth analysis and discussions. Section 5 concludes this paper and provides future directions.

2 Related work

With the development of the Internet, Web services, such as social annotations, have attracted much attention for facilitating users to obtain needed information. Social annotations allow users to assign tags on their interested resources, such as web pages and documents. Therefore, social annotations can not only reflect user preferences, but also capture characteristics of the tagged resources. Complex relationships among users, resources and tags are encoded in social annotations, which can be useful for various web services, such as recommendation system and information retrieval (Godoy and Corbellini 2016; Yu et al. 2018; Mahboob et al. 2017; Xie et al. 2016; Zhou et al. 2017; Wang et al. 2013; Liu et al. 2002; Martin-Bautista et al. 2002; Ibrahim and Landa-Silva 2016; Laura and Me 2017; Abdi et al. 2017). Godoy and Corbellini (2016) provided a comprehensive overview of folksonomy-based recommender systems with a particular focus on link-based retrieval approaches. Yu et al. (2018) presented a novel method to improve the quality of tag recommendation by modeling user preferences. Mahboob et al. (2017) also focused on social annotations-based tag recommendation and proposed a heat diffusion algorithm to tackle the problem. These recommendation-based studies have demonstrated the usefulness of social annotations in user modeling.

To improve the performance of document retrieval, Xie et al. (2016) presented a generic framework to incorporate sentiment information of social tags for personalized search by user profiles and resources profiles. Zhou et al. (2017) proposed a novel model to construct enriched user profiles with social annotations for query expansion. Wang et al. (2013) proposed a novel supervised ranking aggregation method by considering the differences among queries and directly optimizing the NDCG metric. Liu et al. (2002) used content-based technologies to classify and retrieve audio clips based on a fuzzy logic system. Martin-Bautista et al. (2002) investigated user profiles using fuzzy logic in web

retrieval processes. Ibrahim and Landa-Silva (2016) proposed a novel term weighting scheme based on average term occurrences to improve the performance of document retrieval. Laura and Me (2017) summarized the challenges in semantic search engines to help law enforcements against the online drug marketplaces. Abdi et al. (2017) proposed a query-based text summarization method that combines the semantic relations and syntactic compositions among words to reduce the redundancies in summary.

Information retrieval systems aim to search relevant documents or websites for a given query to fulfill user information needs (Lee et al. 2017). Previous studies have sought to integrate social annotations to enhance retrieval performance (Zhou et al. 2008; Hotho et al. 2006; Bouadjenek et al. 2013; Bao et al. 2007; Xu et al. 2008; Du et al. 2016). Bao et al. (2007) incorporated the summary and the popularity of webpages into PageRank algorithm to improve web search. Xu et al. (2008) explored the categorization, keyword and structure of social annotations for personalized search. Du et al. (2016) elaborated on the limitations of the current research on user profiling for Folksonomy-based personalized search and proposed a multilevel user profiling model by integrating tags and ratings to achieve personalized search. These studies have indicated that social annotations are valuable resources to improve information retrieval. We can exploit two kinds of information to boost retrieval performance, including the tags and social structures.

Topic models have been widely used in IR tasks to detect the latent topics of documents. Topic models can stimulate the generation of documents from a probabilistic perspective, such as probabilistic latent semantic indexing (PLSI) model (Hofmann 1999) and the latent Dirichlet allocation (LDA) model (Blei et al. 2003). In particular, the LDA model has been successfully used in text mining tasks based on its solid theoretical foundation and promising performance. A series of variations of LDA have also been proposed to solve various problems (Blei and Jordan 2003; Chen et al. 2009; Erosheva et al. 2004; Liu et al. 2009; Newman et al. 2006; Rosen-Zvi et al. 2004; Lu et al. 2010). For example, Lu et al. (2010) proposed a novel probabilistic generative model to simulate the generation of social annotations for tag prediction. Motivated by their work, we incorporate social annotations into the LDA-based topic models for personalized information retrieval.

Most studies have treated social annotations as complements of queries or documents to capture information of users (Zhou et al. 2008; Ramage et al. 2009). However, the performance has been limited due to the complexity and sparsity of social annotations. How to effectively employ social annotations for information retrieval remains a great challenge. To further improve retrieval performance, we propose to reconstruct candidate documents for retrieval using social annotations and optimize the search process with social annotation-enhanced topic models in this study.

3 Methodology

In this section, we provide more details on our methodology. There are two stages in our method: The first stage reconstructs candidate documents based on social annotations; the second stage optimizes the document-topic distribution produced by latent Dirichlet allocation (LDA) using social information for personalized document retrieval. Based on the reconstructed documents, we incorporate the optimized topic models into query likelihood language model for personalized information retrieval.

3.1 Document reconstruction

Social annotations are valuable resources to measure the quality of documents crawled from websites. When a user annotates a document with certain tags, his or her preferences for the document are embedded in the annotations. Therefore, social annotations can reflect user preferences, which may be helpful in personalized document retrieval. Since social annotations are always sparse, personalized search solely based on social annotations cannot achieve the ideal performance. To fully take advantage of social annotations, we propose to incorporate social annotations into reconstructing the original documents. The reconstructed documents retain unique features of original documents and meanwhile encode user preference based on social annotations. We believe that the reconstructed documents contribute to the improvement of retrieval performance with fewer noisy and redundant contents. Next, we provide details on document reconstruction.

We first represent original documents as feature vectors based on the textual contents and social tags of documents, respectively. These two types of document representations are generated using the vector space model. The dimensionality of the vectors is the vocabulary size of the document collection, and each dimension in these vectors indicates the frequency of each term in documents. Based on these two types of representations, we obtain two representation vectors of each document. Given that users always employs the key words appearing in documents as tags to annotate documents, document representations based on social annotations can enhance the occurrences of key words of documents. We then add these two types of vector representations at the element level to obtain merged document vectors. In the merged document vectors, important words in the original documents are highlighted by social annotations, while other words remain the same. We refine the words in the merged vectors to reduce the noises and redundancy in documents using the term frequency-inverse document frequency (TFIDF) weighting scheme and keep the original number of

words in each document. We formalize this process as follows.

(1) Given a set of documents $D = \{d_1, d_2, \ldots, d_k\}$, all the terms and social tags in the documents constitute a n dimensional vector space. n is the size of vocabulary in the entire document collection. We represent each document d_i as a document representation $W_{d_i} = \{w^1_{d_i}, w^2_{d_i}, \ldots, w^n_{d_i}\}$ and an annotation representation $T_{d_i} = \{t^1_{d_i}, t^2_{d_i}, \ldots, t^n_{d_i}\}$, respectively. Each dimension in the representations is the frequency of each term in the given document. Namely, $w^1_{d_i}$ represents the frequency of the first term in the document representation. $t^1_{d_i}$ represents the frequency of the first term in the annotation representation.
(2) We add these two representations of each document at the element level to obtain a merged document representation $M_{d_i} = \{m^1_{d_i}, m^2_{d_i}, \ldots, m^n_{d_i}\}$ for each document, where $w^1_{d_i}$ represents the merged frequency of the first term. The merged representations of documents highlight the key words in documents, and overshadow the low-frequency words in the original documents with social tags for further optimization.
(3) We adopt the TFIDF weighting scheme to assign weights on each dimension of the merged representations. In the weighting process, term frequency is counted based on the original documents and its corresponding social tags, and inverse document frequency is computed based on the entire document collection. We denote the weights for a given document d_i as $S_{d_i} = \{s^1_{d_i}, s^2_{d_i}, \ldots, s^n_{d_i}\}$. The weights are then combined with the merged representations to obtain a weighted vector for each document, which is denoted as $V_{d_i} = \{v^1_{d_i}, v^2_{d_i}, \ldots, v^n_{d_i}\}$, where $v^k_{d_i} = s^k_{d_i} \times m^k_{d_i}$ for each dimension k.
(4) We sort the words based on the weights in the weighted document vectors and refine the words by choosing the words with the highest weights until the number of words equals to the number of words in each original document. We treat the refined set of words as the reconstructed documents.

The reconstructed documents encode user preferences and capture social information of users from the annotations, which may contribute to personalized document retrieval. We use the reconstructed collection to train the LDA-based topic models with Gibbs sampling to obtain the document-topic distribution and the topic-word distribution.

3.2 Topic model optimization

In this section, we first introduce the latent Dirichlet allocation (LDA) for topic modeling, and then provide our optimization strategy designed for document-topic distribution based on social annotations.

3.2.1 Latent Dirichlet allocation

LDA, as a generative probabilistic topic model, is a widely used technique to identify latent topics of documents within large corpus. The core idea of LDA is that documents are generated as a probabilistic distribution over latent topics, and topics are characterized by a distribution over words. Based on this idea, each document is generated based on the sampling strategy in Table 1.

For each document d, the LDA model generates a probabilistic distribution θ over topics from the prior Dirichlet distribution α. Each word w_m in the document d is generated based on the topic z_m and the Dirichlet prior β. We estimate the joint probability of a topic mixture θ, a set of topics $\mathbf{z}$ and a set of M words $\mathbf{w}$ as follows.

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{m=1}^{M} p(z_m|\theta) p(w_m|z_m, \beta) \tag{1}$$

We then obtain the marginal distribution of a document by integrating over θ and summing over z.

$$p(w|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{m=1}^{M} p(z_m|\theta) p(w_m|z_m, \beta) \right) \mathrm{d}\theta \tag{2}$$

We finally obtain the probability of the entire document collection by taking the production of the marginal probabilities of single documents.

$$p(D|\alpha, \beta) = \prod_{d=1}^{|D|} \int p(\theta_d|\alpha) \left(\prod_{m=1}^{M_d} p(z_{dn}|\theta) p(w_m|z_{dn}, \beta) \right) \mathrm{d}\theta_d \tag{3}$$

The outputs of LDA are two distributions: the document-topic distribution and the topic-word distribution. Different methods have been used to estimate these distributions.

Table 1 Latent Dirichlet allocation

1. Choose N ~Poisson(ξ)
2. Choose θ ~Dirichlet(α)
3. For each of the N words w_m:
   (a) Choose a topic $z_m \sim$ Multinomial(θ)
   (b) Choose a word w_m from $p(w_m|z_m, \beta)$
   a multinomial probability conditioned on the topic z_m

Topic distributions depict the various aspects of documents with respect to given queries and contribute to personalize retrieval results for more preferable documents. To make the most of topic information, we modify the topic distribution using social annotations to obtain enhanced topic model for advanced retrieval.

3.2.2 Social annotation-enhanced topic model

In our method, we learn the LDA-based topic model using the reconstructed document collections and obtain the document-topic distribution and the topic-word distribution. To further take advantage of the social information of users, we modify the topic distribution with social annotations.

In social annotations, each document d is annotated by a set of users U_d. Each user u_i annotates the document d with a set of tags $Tag_d^{u_i}$. Therefore, users and documents are associated with social tags, indicating user interests on the documents in certain topics. To this end, we attempt to represent the documents with user tags as complements to enhance learned topic models.

Specifically, LDA assumes that each word belongs to each topic with a certain probability, and each topic contains each word with a certain probability. Each tag in $Tag_d^{u_i}$ reflects the interests of u_i on the document and can be correlated with a certain probability. We therefore map the tags with the probability that the tag words belongs to each topic and sum all the probabilities of all the tags in $Tag_d^{u_i}$ to obtain a vector $G_d^{u_i}$. $G_d^{u_i}$ is a non-normalized probability of topic distribution, capturing user preferences from social annotations. We then sum all the vectors at the element level for all the annotated users of the document d and obtain a new document-topic probability distribution G^d. We finally integrate the new probability distribution into the learned document-topic distribution obtained from LDA as follows.

$$\theta_{\text{new}}^d = \gamma\theta_{\text{old}}^d + (1-\gamma)G^d \tag{4}$$

where θ_{new}^d is the new document-topic distribution for each document d, and θ_{old}^d is the LDA-based document-topic distribution for any document d. γ is the hyper-parameter for linear interpolation, which controls the weights on social annotation-based topic distribution. When γ equals 0, the new probability distribution will be determined based on social annotations. When γ equals 1, the new probability distribution will be determined based on the original LDA-based topic distribution. We normalize the probability distribution for further optimization in personalized document retrieval. The original LDA-based distribution models the topics of documents based only on the contents of documents, which ignores the user preference information for personalized document retrieval. To capture the user preferences, we combine the LDA-based topic distribution with the social annotation-based topic distribution in Eq. (4). In the new document-topic distribution, the probability of the modified topics magnifies the likelihood of retrieving personalized results using annotation information, which may promote the ranking of documents in consideration of user preferences.

3.3 Personalized document retrieval

We adopt the query likelihood language model for personalized document retrieval. The query likelihood language model has been widely used in information retrieval tasks, particularly on personalized document retrieval (Zhou et al. 2008). This model, as an important probabilistic retrieval model, has been proved effective in improving the retrieval and ranking accuracy of documents. Since the model is general, abundant external information, such as user preferences, can be well integrated for task-specific retrieval performance. The model assumes that query terms are generated by a probabilistic model based on certain observed documents. Formally, language models can be estimated as follows.

$$P(Q|d) = \prod_{q\in Q} P(q|d) \tag{5}$$

where q represents a query term in a given query Q. d is an observed document. Candidate documents for the query Q are then ranked based on their probabilities as the final ranking list of documents.

The existing studies have demonstrated that smoothing methods are effective in improving the performance of language models. Different smoothing methods have been investigated in IR tasks, among which the Jelinek–Mercer method is particularly effective in the LDA-based retrieval. The Jelinek–Mercer method involves a linear interpolation of the document language model and the collection model, which can be formalized as follows.

$$\begin{aligned} P(q|d) = \tau\left(\frac{N}{N+\delta}P_d(q|d) + \left(1-\frac{N}{N+\delta}\right)P(q|\text{coll})\right) \\ +(1-\tau)P_s(q|d) \end{aligned} \tag{6}$$

where $P_d(q|d)$ represents the maximum likelihood of a query term q in the document d. $P(q|\text{coll})$ represents the maximum likelihood of the term q in the entire document collection. N is the total number of documents in the entire collection. δ and τ are tunable parameters in the method. In the original method, $P_s(q|d)$ is based on the document-topic distribution obtained from LDA. In our method, we attempt to incorporate user preferences from social annotations into the distribution. The modified $P_s(q|d)$ will contribute to improving personalized document retrieval. Therefore, we compute $P_s(q|d)$ using the modified social annotation-enhanced topic model θ_{new}^d based on Eq. (4). The modified $P_s(q|d)$ can be formalized as follows.

$$P_s(q|d) = \sum_{i}^{K} P(q|\beta)\frac{\phi_{d_i}}{\sum_{j=1}^{K}\phi_{d_j}} \quad (7)$$

where $P(q|\beta)$ represents the probability that a query term q belongs to certain topics. ϕ_{d_j} represents the probability that the document d contains the topic i based on the social annotation-enhanced document-topic distribution.

Overall, our method incorporates user preferences from social annotations into query likelihood language model with Jelinek–Mercer smoothing. The advantages of the proposed method are twofold. On the one hand, we reconstruct the candidate documents with more emphasis on the user-centered key words and less on the redundant words. On the other hand, we modify the LDA-based topic distributions using social annotations and integrate the new distribution into the query likelihood language model to capture user preferences. Therefore, we believe that our method can enhance personalized document retrieval based on social annotations.

4 Experiments and analysis

In this section, we evaluate the proposed method using extensive experiments and provide in-depth analysis and discussions on the experimental results.

4.1 Experimental settings

We use the dataset from Del.ici.ous[1] to evaluate the proposed method in our experiments. Del.ici.ous is one of the largest social annotation websites. The dataset we used has been used in the previous study (Lu et al. 2010), which contains 4414 users, 41,190 documents and 28,740 tags. Each document has been annotated with at least 20 social tags. The number of tags for each document is at least 1.12% of the number of words in the document, and more than 60% documents have more than 5% tag words. Statistics of the dataset implies that the dataset is suitable to examine the effectiveness of the proposed method. We treat the 50 most frequent tags in the dataset as queries and report the retrieval performance based on the average performance of all the queries.

We preprocess the data by removing the stopwords and stemming the words for accurately matching the words. To train the LDA model, we use Gibbs sampling to capture the latent topics of the document collection D. We empirically set the hyper-parameters $\alpha = 0.625$, $\beta = 0.1$ and $num_{\text{topics}} = 80$ following the previous study (Zhou et al. 2008; Lu et al. 2010), since the optimal performance can be achieved using these settings. We tune the parameter γ on a development set and observe that relatively good performance can be achieved when set $\gamma = 0.8$. We use 10% of the data as a development set for parameter tuning in our experiments.

Since there is no publicly available dataset for social annotation-based information retrieval, we manually evaluate the quality of the ranking list of documents by ten graduate students and report the average performance. Each student annotates the relevance between a query and a retrieved document with labels ranging from 1 to 5, where label 1 indicates the irrelevance of a document and label 5 indicates that a document is the most relevant one.

To ensure the correctness of the annotations, annotators were given standards detailed instruction for potential difficult and common problems with many annotated samples. Aside from giving them the detailed guidelines, we gave them a formal training lesson and a laboratory meeting to exchange ideas and to discuss problems about annotation. These ten annotators were divided to four groups with three members for each group plus one group with one person. Using cross-validation methods for annotation, the three-member groups annotated, and the one-person group participated in the final decision when there was divergence. If there was no divergence between members in the same group, the annotating work was complete. Otherwise another group annotated again, and the final group annotated if there was still divergence. Finally, if the groups could not reach agreement on the annotation, everyone discussed and determined the annotation to ensure its accuracy and consistency.

We make public the ranked dataset and the source code for easily reproducing and comparing in relevant future studies.[2] We evaluate the performance in terms of classic IR evaluation metrics $P@5$, $P@10$, $NDCG@5$, $NDCG@10$ and $MAP@10$. These metrics can completely evaluate the performance, and higher values indicate higher retrieval performance. We average the metric values by each annotator to obtain an average performance for fair comparisons.

4.2 Compared models

We compare the proposed method with four state-of-the-art models in our experiments. In this section, we introduce these models together with different versions of the proposed model for examining the effectiveness of the proposed optimization strategies.

The first baseline model is WT-QDAU (Zhou et al. 2008). The model used social annotations for personalized information retrieval with a computationally tractable hierarchical Bayesian network to combine term-level language models with topic-level models obtained from topics in documents and users. The second baseline model is FolkRank (Hotho et al. 2006). The model is learned using a new search algorithm FolkRank based on folksonomies, which used the

[1] https://del.icio.us/.

[2] https://www.jianguoyun.com/p/DZWqIdgQ_66nBxiThagB.

structure of folksonomy to enhance retrieval performance. The third baseline model is QE-SA (Zhou et al. 2017). The model constructed enriched user profiles with the help of an external corpus for personalized query expansion and integrated word embeddings with topic models in two groups of pseudo-aligned documents. The fourth baseline model is PSDR (Bouadjenek et al. 2013). The model involved a personalized document representation approach for document ranking based on social activities of users. Parameters of these baseline models are tuned using a development set, and we report the optimal performance in the comparisons.

For the proposed method, we report the results of LDA-enhanced language model as the basic model and then modify the model with document reconstruction (LDA-DR) and topic optimization (LDA-TO), respectively, to examine the effectiveness of these two optimization strategies. We finally report the results of the proposed model (LDA-DR-TO). We also report the experimental results of our model with different parameters to illustrate the parameter tuning. The chosen baselines, together with the proposed methods, are under the same experimental settings for fair comparison. Therefore, we believe that the comparisons in our experiments are reliable to demonstrate the effectiveness of the proposed methods.

4.3 Experimental results

In this section, we report the experimental results of different models in Table 2. The performance is averaged over all the fifty queries. Two-tailed paired Student t-tests ($p < 0.05$) are used to examine whether the improvements are significant relative to the baseline models, where an asterisk indicates significant improvements over the QE-SA model and a dagger indicates significant improvements over the PSDR model.

From the results, we observe that different baseline models achieve diverse ranking performance, and the PSDR model achieves the best performance among all the baseline models. The QE-SA model achieves comparable results to the PSDR model. The LDA model outperforms the WT-QDAU and FolkRank model, but yields slightly worse performance than the other two baseline models. The LDA model with document reconstruction outperforms the QE-SA and PSDR model in terms of most evaluation metrics. Similar trend can be observed on the results of the LDA model with topic optimization. This finding implies that document reconstruction and topic optimization can both contribute to retrieval performance, and social annotations are useful for personalized retrieval of documents. Furthermore, we observe that our model with both of the optimization strategies achieves the best performance among all the compared model. This indicates the effectiveness of the proposed models.

To further examine the robustness of our model with different parameters, we illustrate the retrieval performance of our models by switching the parameter γ and τ. γ controls the interpolation ratio of the LDA-based topic model and the annotation-based topic model, and τ controls the weights of the topic model in the LDA-enhanced language model.

We report the experimental results with different values of the parameter γ in Fig. 1, in which we compare four variations of our methods. The figure indicates that our method with document reconstruction and topic optimization outperforms other models with different γ values. The best performance can be achieved when setting $\gamma = 0.8$. When γ is set as values less than 0.5, the retrieval performance varies and fluctuates for different models and monotonically increases with the increase in the value. When γ is set more than 0.5, the retrieval performance becomes more steady and slightly increases with larger values. Since γ controls the impact the social annotations in our methods, the results indicate that LDA-based topic distribution plays an important role in relevance matching, and social annotations contribute more information on user preferences to enhance the personalized retrieval performance.

We report the experimental results with different values of the parameter τ in Fig. 2. Different values of the parameter τ can yield consistent improvement by our method, and the best performance can be obtained when setting $\tau = 0.7$. We achieve significant improvement over the best performance of the baseline models. This finding also shows the robustness of the proposed model. Since τ reflects the impact of

Table 2 Result comparison of different models

Models	NDCG@5	NDCG@10	P@5	P@10	MAP
WT-QDAU	0.6213	0.6538	0.6335	0.6661	0.4332
FolkRank	0.6265	0.6567	0.6326	0.6662	0.4358
QE-SA	0.6449	0.6631	0.6433	0.6791	0.4413
PSDR	0.6674	0.6825	0.6689	0.6911	0.4523
LDA	0.6314	0.6598	0.6329	0.6677	0.4389
LDA-DR	0.6615∗	0.6853 ∗ †	0.6611∗	0.7018 ∗ †	0.4598 ∗ †
LDA-TO	0.6760 ∗ †	0.6938 ∗ †	0.6679∗	0.7072 ∗ †	0.4617 ∗ †
LDA-DR-TO	0.6823 ∗ †	0.7014 ∗ †	0.6735 ∗ †	0.7298 ∗ †	0.4662 ∗ †

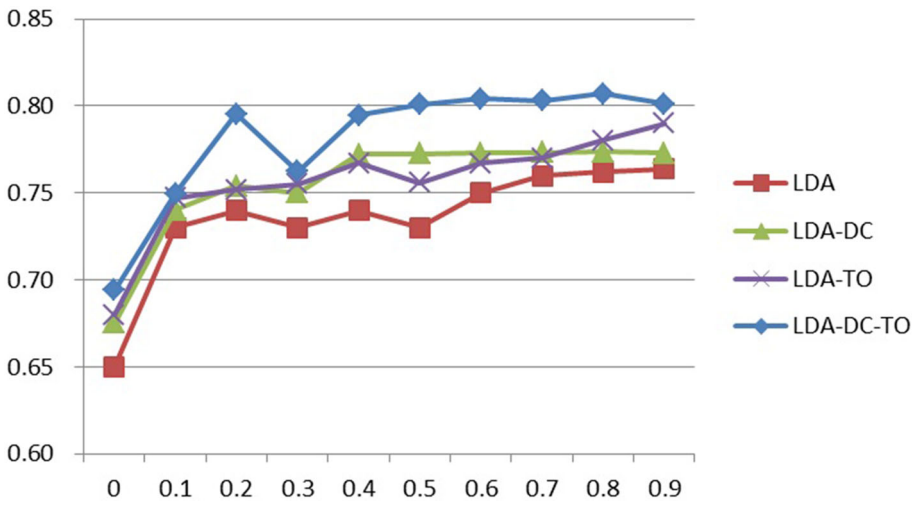

Fig. 1 Evaluation on performance with different values of the parameter γ

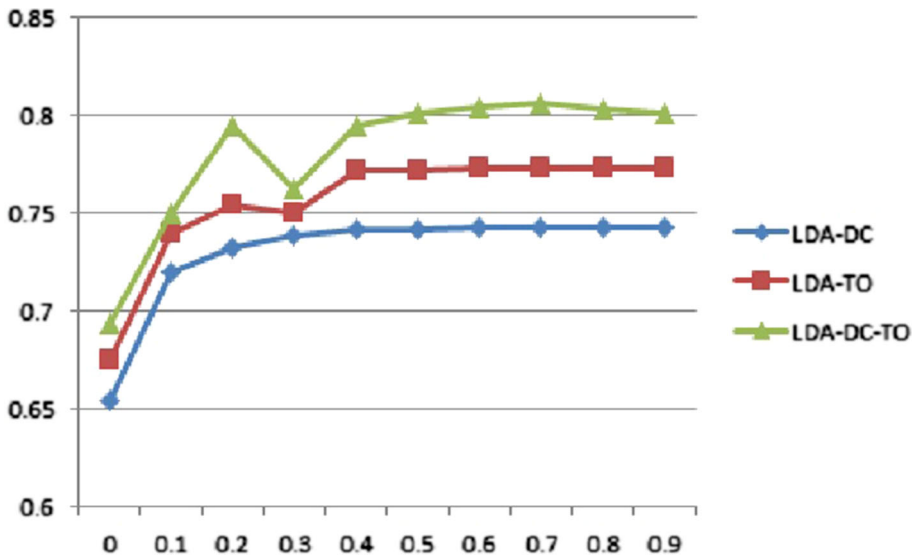

Fig. 2 Evaluation on performance with different values of the parameter τ

social annotation-based topic distribution in the query likelihood language model, we observe that a larger value of τ contributes to the improvement of retrieval performance. This finding is consistent with the results in Fig. 1. Therefore, we believe that social annotations help incorporate user preferences through LDA-based topic models to enhance the personalized retrieval performance.

Furthermore, we investigate the influence of the number of queries on personalized retrieval performance. In our evaluation, we treat the most frequent tags as queries. We switch the number of the most frequent tags from 10 to 50 and report the average retrieval performance with respect to different numbers of queries in Fig. 3. The queries are sorted by their frequencies from high to low in advance. From the results, we observe that our method produces the highest performance with the most frequent 10 tags as queries. This is because the most frequent tags involve more user information, and therefore, the user preferences can be better modeled for these queries. The results indicate that social annotation is effective in improving personalized retrieval performance. Abundant social annotations can contribute more to the improvement of the performance.

We attribute the improvement of our method in two respects. One is document reconstruction, which highlights the important words in documents by considering user tags and meanwhile removes the redundancies of documents by limiting the length of the reconstructed documents. The other is topic optimization. Topic optimization enhances the topic model learned using LDA with social annotations. Social annotations comprehensively capture the complicated relationship among users, documents and tags, and contribute to more effective topic models for document ranking.

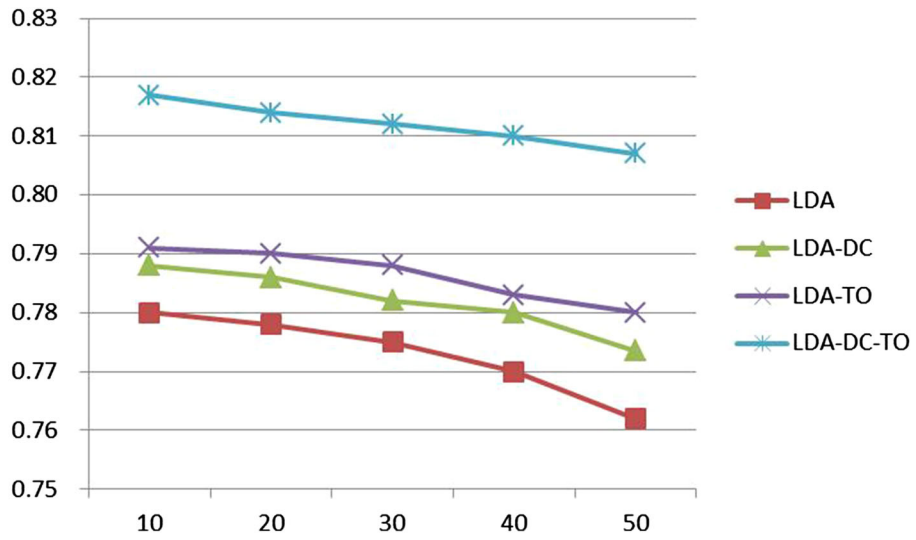

Fig. 3 Evaluation on performance with different numbers of the most frequent tags as queries

5 Conclusions and future work

In this paper, we propose a novel method based on social annotations for personalized document retrieval. Our method first reconstructs candidate documents using social tags for documents to capture user preferences, and then, we generalize the LDA-based topic models by considering the relationship among users, social tags and documents from social annotations. Our method therefore encodes comprehensive information of social users into the ranking process of documents, thus largely enhancing personalized document retrieval. Experimental results demonstrate the effectiveness of our method in comparison with baseline methods. Our future work can be carried out from two directions. One is to integrate the social annotation-enhanced topic models into other types of topic models for accurately capturing the user-oriented latent topics of documents. The other is to optimize the ranking models, such as learning to rank models, based on social annotations for more useful ranking lists of documents.

Acknowledgements This work is partially supported by Grant from the Natural Science Foundation of China (Nos. 61632011, 61572102, 61602078, 61572098), the Ministry of Education Humanities and Social Science Project (No. 19YJCZH199), the China Postdoctoral Science Foundation (No. 2018M641691), the Fundamental Research Funds for the Central Universities (No. DUT18ZD102) and the National Key Research Development Program of China (No. 2016YFB1001103).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Human and Animal Rights This article does not contain any studies with human participants or animals performed by any of the authors.

References

Abdi A, Idris N, Alguliyev RM, Aliguliyev RM (2017) Query-based multi-documents summarization using linguistic knowledge and content word expansion. Soft Comput 21(7):1785–1801

Bao S, Xue G, Wu X, Yu Y, Fei B, Su Z (2007) Optimizing web search using social annotations. In: Proceedings of the 16th international conference on World Wide Web. ACM, pp 501–510

Blei DM, Jordan MI (2003) Modeling annotated data. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 127–134

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3(Jan):993–1022

Bouadjenek MR, Hacid H, Bouzeghoub M, Vakali A (2013) Using social annotations to enhance document representation for personalized search. In: Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval. ACM, pp 1049–1052

Chen X, Lu C, An Y, Achananuparp P (2009) Probabilistic models for topic learning from images and captions in online biomedical literatures. In: Proceedings of the 18th ACM conference on information and knowledge management. ACM, pp 495–504

Du Q, Xie H, Cai Y, Leung H, Li Q, Min H, Wang FL (2016) Folksonomy-based personalized search by hybrid user profiles in multiple levels. Neurocomputing 204:142–152

Erosheva E, Fienberg S, Lafferty J (2004) Mixed-membership models of scientific publications. Proc Natl Acad Sci 101(suppl 1):5220–5227

Godoy D, Corbellini A (2016) Folksonomy-based recommender systems: a state-of-the-art review. Int J Intell Syst 31(4):314–346

Golder SA, Huberman BA (2006) Usage patterns of collaborative tagging systems. J Inf Sci 32(2):198–208

Hofmann T (1999) Probabilistic latent semantic analysis. In: Proceedings of the 15th conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., pp 289–296

Hotho A, Jäschke R, Schmitz C, Stumme G (2006) Information retrieval in folksonomies: search and ranking. In European semantic web conference. Springer, pp 411–426

Ibrahim OAS, Landa-Silva D (2016) Term frequency with average term occurrences for textual information retrieval. Soft Comput 20(8):3045–3061

Laura L, Me G (2017) Searching the web for illegal content: the anatomy of a semantic search engine. Soft Comput 21(5):1245–1252

Lee S, Masoud M, Balaji J, Belkasim S, Sunderraman R, Moon S-J (2017) A survey of tag-based information retrieval. Int J Multimed Inf Retr 6(2):99–113

Lin Y, Lin H, Jin S, Ye Z (2011) Social annotation in query expansion: a machine learning approach. In: Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval. ACM, pp 405–414

Liu M, Wan C, Wang L (2002) Content-based audio classification and retrieval using a fuzzy logic system: towards multimedia search engines. Soft Comput 6(5):357–364

Liu Y, Niculescu-Mizil A, Gryc W (2009) Topic-link lda: joint models of topic and author community. In: Proceedings of the 26th annual international conference on machine learning. ACM, pp 665–672

Lu C, Hu X, Chen X, Park J-R, He TT, Li Z (2010) The topic-perspective model for social tagging systems. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 683–692

Mahboob VA, Jalali M, Jahan MV, Barekati P (2017) Swallow: resource and tag recommender system based on heat diffusion algorithm in social annotation systems. Comput Intell 33(1):99–118

Martin-Bautista MJ, Kraft DH, Vila MA, Chen J, Cruz J (2002) User profiles and fuzzy logic for web retrieval issues. Soft Comput 6(5):365–372

Newman D, Chemudugunta C, Smyth P (2006) Statistical entity-topic models. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 680–686

Pantel P, Gamon M, Alonso O, Haas K (2012) Social annotations: utility and prediction modeling. In: Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval. ACM, pp 285–294

Ramage D, Heymann P, Manning CD, Garcia-Molina H (2009) Clustering the tagged web. In: Proceedings of the 2nd ACM international conference on web search and data mining. ACM, pp 54–63

Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P (2004) The author-topic model for authors and documents. In: Proceedings of the 20th conference on uncertainty in artificial intelligence. AUAI Press, pp 487–494

Wang Y, Huang Y, Pang X, Lu M, Xie M, Liu J (2013) Supervised rank aggregation based on query similarity for document retrieval. Soft Comput 17(3):421–429

Wu X, Zhang L, Yu Y (2006) Exploring social annotations for the semantic web. In: Proceedings of the 15th international conference on World Wide Web. ACM, pp 417–426

Xie H, Li X, Wang T, Lau RYK, Wong T-L, Chen L, Wang FL, Li Q (2016) Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy. Inf Process Manag 52(1):61–72

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

Yu H, Zhou B, Deng M, Hu F (2018) Tag recommendation method in folksonomy based on user tagging status. J Intell Inf Syst 50(3):479–500

Zhou D, Bian J, Zheng S, Zha H, Giles CL (2008) Exploring social annotations for information retrieval. In: Proceedings of the 17th international conference on World Wide Web. ACM, pp 715–724

Zhou D, Wu X, Zhao W, Lawless S, Liu J (2017) Query expansion with enriched user profiles for personalized search utilizing folksonomy data. IEEE Trans Knowl Data Eng 29(7):1536–1548