Research on Hybrid Tag Recommendation Based on LDA Topic Model (Postprint)

Authors: Xiong Huixiang, Dou Yan

Date: 2023-08-26T00:00:00+00:00

Abstract

[Purpose/Significance] To address the problem of unsatisfactory results obtained by current tag recommendation methods, this study improves traditional similarity computation approaches and integrates multiple tag recommendation methods to enhance recommendation accuracy. [Method/Process] By fusing content-based and collaborative filtering recommendation concepts, LDA is employed for similarity computation to derive neighbor sets of resources and users, and resource content keywords are extracted to construct a hybrid tag recommendation model. The model is validated using "Douban Reading" as a case study and compared with several existing tag recommendation methods. [Results/Conclusion] In social tagging systems, it is imperative to consider the three dimensions of user-resource-tag; considering only a single perspective inevitably leads to incomplete results. Moreover, introducing LDA into similarity computation can uncover latent semantic relationships and improve recommendation quality, while combining multiple methods to leverage their respective strengths can yield more satisfactory recommendation results.

Full Text

Preamble

Vol. 62 No. 3 February 2018 ChinaXiv Cooperative Journal

Research on Tag Hybrid Recommendation Based on LDA Topic Model

Xiong Huixiang, Dou Yan
School of Information Management, Central China Normal University, Wuhan 430079

Abstract

[**Purpose/Significance**] Current tag recommendation methods produce unsatisfactory results. This study improves traditional similarity calculation methods and combines multiple tag recommendation approaches to enhance recommendation accuracy.

[**Method/Process**] Integrating content-based and collaborative filtering recommendation ideas, we use LDA for similarity calculation to obtain neighbor sets of resources and users, extract keywords from resource content, and construct a tag hybrid recommendation model. Using "Douban Reading" as a case study, we validate the model and compare it with several tag recommendation methods.

[**Result/Conclusion**] In social tagging systems, three dimensions—user, resource, and tag—must be considered. Considering only a single perspective inevitably leads to incomplete results. Introducing LDA in similarity calculation can mine potential semantic relationships, improve recommendation quality, and combining multiple methods can produce more satisfactory recommendation results.

Keywords: social tagging; tag recommendation; collaborative filtering; LDA
Classification Number: TP181
DOI: 10.13266/j.issn.0252-3116.2018.03.013

Social tagging is a primary and effective method for organizing network information resources in the Web 2.0 era, allowing users to annotate various online resources with custom keywords (tags) for effective organization, retrieval, and utilization. Tags, created by users without restrictions, reflect users' understanding of resources and enable users to retrieve resources or find others with similar interests through tags. While widely applied on the Internet, social tagging has generated many problems. For example, due to differences in user preferences, different users may use different tags to annotate the same resource. The uncontrolled nature of social tags leads users to create erroneous tags or meaningless garbage tags during free annotation. Using a single recommendation method results in one-sided outcomes. Additionally, these traditional recommendation techniques fail to consider the rich semantic information contained among tags or preference differences arising from various factors, leading to inadequate recommendation quality. Therefore, this paper proposes a hybrid tag recommendation method combining multiple technologies, integrating three approaches: content-based tag recommendation, user-based collaborative filtering, and resource-based collaborative filtering. It introduces Latent Dirichlet Allocation (LDA) into the similarity calculation process, using topic probability distributions as the basis for calculation instead of traditional methods, incorporating deep semantic knowledge to generate recommendation tags based on similar resources and similar users. Second, it extracts keywords from resource content as content-based recommendation tags. Finally, it fuses the three results to recommend tags to users. The significance lies in addressing three tag

source perspectives, fusing multiple recommendation technologies to correspond recommendation methods with tag source angles, improving data density, and thus avoiding the defects of single technologies. Experiments show that this LDA-based hybrid tag recommendation method alleviates tag semantic ambiguity to some extent and significantly improves recommendation results.

Current domestic and international research on social tagging mainly focuses on resource recommendation and user recommendation, with fewer studies on tag recommendation. Existing research is limited to single technologies such as content-based tag recommendation, collaborative filtering-based tag recommendation, or association rule-based tag recommendation. There are few studies combining multiple technologies. Tags in tagging systems have multiple sources, and using only a single technology has unavoidable defects.

2 Tag Recommendation and Related Technologies

2.1 Tag Recommendation

Social tags represent user interest preferences and can reveal resource characteristics from different dimensions. However, their excessive freedom leads to many low-quality tags in the system, affecting the effectiveness of tags to some extent. The most direct method to improve tag quality is to control tag usage, but this compulsory and restrictive approach is difficult for users to accept and contradicts the freedom of social tagging. Therefore, tag recommendation mechanisms emerge. Tag recommendation refers to recommending a series of relevant tags to users when they want to annotate a resource, based on their tagging history, resource content features, and existing tags in the system. As an auxiliary tool for user annotation, tag recommendation can provide references and suggestions, reduce user burden, enhance tagging enthusiasm, improve tag quality, and increase resource retrieval efficiency and accuracy. Compared with forced control of tag usage, using the friendlier and gentler suggestive interaction language of tag recommendation to appropriately standardize user tagging behavior simplifies the annotation process, improves user experience, and enhances annotation quality, making it highly significant for research.

2.2 Personalized Recommendation Technology

Current tag recommendation methods in social tagging systems actually borrow from personalized recommendation technologies in e-commerce. Current domestic personalized recommendation technologies mainly include three types: content-based recommendation, collaborative filtering recommendation, and hybrid recommendation. Content-based recommendation relies on resource content itself, typically using text content. Collaborative filtering includes resource-based collaborative filtering and user-based collaborative filtering, calculating similarity between users or resources for recommendation. Since each recommendation technology has its own advantages and disadvantages, hybrid recommendation has been frequently proposed in recent years, with the complementarity

of multiple recommendation methods helping to avoid the shortcomings of single methods and improve recommendation quality. With technological advances, new methods have emerged, such as association rule-based recommendation, link prediction-based recommendation, and social network trust relationship-based recommendation.

Many scholars have proposed various tag recommendation techniques based on the above recommendation technologies combined with tag characteristics. Foreign scholars M. Tatu et al. combined nearest neighbor methods with keyword extraction for word-based tag recommendation. G. Mishne used K-nearest neighbor methods to select K resources most relevant to the resource to be annotated from document collections and recommended their tags to users. L. Marinho et al. used the relationships among tag-user-resource to obtain two-dimensional matrices between each pair, discovering similar users through matrix vectors for tag recommendation. A. Hotho et al. proposed the FolkRank algorithm based on PageRank for recommendation using link analysis. In China, Song Hongxin et al. conducted research on Sina blog tags, proposing a recommendation model based on keyword extraction and blog article classification. Gao Bing focused on tag recommendation technology in Q&A communities, recommending tags based on finding similar annotated questions. Wang Chuanbao proposed a hybrid tag recommendation method based on collaborative filtering and text similarity. An Zhiwei proposed a tag recommendation algorithm based on three-part graph tensor decomposition. Zhang Liang fused the relationships among user, tag, and resource, directly using LDA to build a unified topic model for tag recommendation.

The analysis shows that: (1) Most recommendation technologies in tag recommendation systems are content-based, collaborative filtering-based, or other single-angle methods, with few hybrid methods; (2) Collaborative filtering, tensor decomposition, and other methods analyze relationships among user-resource-tag, ignoring tag semantics and resource content features; (3) LDA applications in tag recommendation mainly focus on using LDA for tag clustering, semantic analysis, or direct recommendation, rarely combining LDA with other tag recommendation methods. Therefore, this paper addresses these problems from multiple perspectives, combining currently popular recommendation technologies, considering both inter-object relationships and tag semantic content, and using LDA to improve traditional similarity calculation for tag recommendation.

2.3 LDA Topic Model

LDA is an unsupervised probabilistic topic model commonly used for modeling large-scale document collections. It is based on the assumption that when writing a document, a user must have certain determined topics in mind. With topics determined, the user will select words from a word pool of a certain topic with a certain probability to illustrate the topic, with the entire document equivalent to a mixture of different topics. The core idea of LDA is shown in Formula

(1):

p(词语　文档)= 主题 (词语 | 主题) × p(主题　公式 (1)

LDA is essentially a three-layer Bayesian probability model with a document-topic-vocabulary hierarchical structure, as shown in Figure 1 [Figure 1: see original paper]:

Figure 1 LDA Model

It uses Dirichlet distribution as the prior distribution for the multinomial distribution of the probabilistic topic model. In this model, W represents vocabulary, the only observable variable; M represents the entire document collection; N represents the total number of words in each document; K represents the number of topics; α and β represent the hyperparameters of the prior distributions for document-topic probability and topic-word probability distribution ϕ, respectively. LDA adopts the bag-of-words idea, first selecting a topic with a certain probability, then selecting a word under that topic with a certain probability, repeating this process until all words in the document are generated. This recommendation technology indirectly performs fuzzy clustering on vocabulary. By training to obtain the distribution of each document on topics and the distribution of each topic on the word space, it mines text information, measures potential semantic relationships among documents, has powerful dimensionality reduction capabilities, and alleviates data sparsity problems. Therefore, introducing the LDA topic model in similarity calculation can measure similarity well even when users use different tags or resources are represented by different feature words, as long as these words belong to the same topic, thereby improving recommendation quality in sparse environments.

3 Recommendation Framework Description and Data Preprocessing

3.1 Tag Recommendation Model Description

Although current tag recommendation methods are diverse, each has unavoidable drawbacks. Moreover, tag recommendation systems differ from ordinary e-commerce recommendation systems as they contain three elements: user-tag-resource. Tags, as intermediaries connecting users and resources, include three tag sources: resource content tags, resource popular tags, and user interest tags, each with different characteristics. Recommending from only one perspective leads to missing data and inevitably causes biased and incomplete results. Therefore, to improve tag recommendation accuracy by fusing multiple tag sources, this paper proposes a recommendation model (see Figure 2) with hybrid recommendation as the breakthrough point. The fusion principle lies in first using different recommendation techniques according to different tag sources—content-based recommendation from the resource content tag perspective, resource-based collaborative filtering from the resource popular tag perspective, and user-based collaborative filtering from the user interest tag

perspective—then fusing the results of these three recommendation techniques. The final recommendation result must contain all tag sources, covering user-tag-resource three aspects. This method combines inter-object relationship analysis with tag semantic analysis, coarsening data granularity to make it denser, thus avoiding the shortcomings of single methods.

The overall recommendation framework is shown in Figure 2 [Figure 2: see original paper], consisting of five stages: data collection, data preprocessing, LDA training, similarity calculation, and recommendation result generation.

In this model, assuming there are m users in user set U, n books in book resource set R, and p tags in tag set T, when user u U annotates resource r R, the system establishes a resource-word corpus and a user-tag corpus, uses LDA to obtain the probability distribution of resources and users on topics for similarity calculation, finds the neighbor user set of user u and neighbor resource set of resource r, and finally combines recommendation tags obtained from similar resources and similar users with keywords extracted from resources to recommend relevant tags t T to user u. Since the data volume in this paper is small, related settings such as k values are not discussed. The focus is on clearly explaining the construction and implementation of the tag recommendation model. According to domestic and foreign literature, general LDA model parameters are set as α=50/k (k is the number of topics), β=0.01. This paper uses these parameters as the benchmark for modeling.

3.2 Experimental Data

3.2.1 Data Collection Douban is a popular social tagging website in China where users can annotate resources including books, music, and movies. Users can browse resources with the same tag to find new interests or find users who have annotated the same resources. Therefore, this paper takes Douban's "Douban Reading" channel as the research object to explain the tag recommendation model. We randomly selected 25 users from "Douban Reading" through manual browsing and collected data from their "Reading," "Want to Read," and "Read" sections, including book titles, book introductions, common tags for each book (10 tags), and tags each user used to annotate relevant books as the experimental research foundation.

3.2.2 Data Preprocessing First, we used the NLPIR Chinese word segmentation system from the Chinese Academy of Sciences to segment book introductions and non-standard tags, and used a stop word list to filter out meaningless words (such as "ah," "just," "hey," etc.) and special symbols. Meanwhile, tags containing English were all converted to lowercase. Additionally, proper nouns such as book titles and author names are important parts describing resource features and key sources for tag recommendation. Therefore, we used the user-defined dictionary function of NLPIR to add these words to the custom dictionary for retention during word segmentation. After processing, we obtained: 25 users, 135 book resources with their introductions, 592 common

book tags, and 234 user tags, as shown in Tables 1 and 2 .

4 Tag Hybrid Recommendation Based on LDA

4.1 Resource Topic Model Training and Calculation

4.1.1 Resource-Topic Model Training After data preprocessing in Section 3.2, any book resource r is represented by n introduction words w, as shown in Table 3 . Treating books as documents and introduction words as words in documents, we use LDA for modeling to obtain resource-topic probability distributions and topic-word probability distributions. Subsequent calculations only need to use the probability distribution of resources on topics. Using Python and its LDA toolkit to train Table 3 with topic number k=15, we obtain the resource-topic probability distribution, i.e., the probability of each topic appearing in each resource, as shown in Table 4 .

4.1.2 Similarity Calculation

(1) Topic Probability Distance. Traditional collaborative filtering technologies often use cosine similarity or Pearson correlation coefficient to calculate similarity. However, since this paper's calculation is based on resource distribution on topic probabilities, traditional formulas cannot be used directly. KL (Kullback-Leibler) divergence, also known as KL distance, is commonly used to calculate the distance between two probability distributions, as shown in Formula (2):

Dkl(p,q)= i=1 pi ln 公式 (2)

In Formula (2), p and q are two probability distributions, and for any i, when pi=qi, Dkl(p,q)=0. However, KL divergence is an asymmetric distance. Therefore, for convenience, its symmetric formula JS (Jensen-Shannon) divergence is often used, as shown in Formula (3):

Djs(p,q)= [Dkl(p,)+Dkl(q,)] 公式 (3)

In Formula (3), p and q are also two probability distributions. This formula's range is [0,1], meaning the smaller the JS divergence value, the closer the distance between the two probabilities; the larger the value, the farther the distance. This paper selects this distance formula to calculate the distance of topic probability distribution between two book resources based on Table 4, with results shown in Table 5 .

(2) Matrix Conversion. Table 5 shows the distance differences between book resources. According to Formula (3), smaller values indicate closer distance between two individuals. For convenience in subsequent calculations, we convert it using Formula (4):

Sim(a,b)=1/(1+D(a,b)) 公式 (4)

In Formula (4), Sim(a,b) is the similarity between book resources a and b, and D(a,b) is the topic distribution distance between a and b. Adding 1 to

the denominator prevents the impact when distance is 0. A larger Sim value indicates greater similarity. Calculation results using Formula (4) are shown in Table 6 .

4.2 User Topic Model Training and Calculation

4.2.1 User-Topic Model Training Treating users as documents and tags used by users as words in documents, as shown in Table 7 , we use Python for LDA modeling of Table 7 with topic number k=5. After training, we obtain the user-topic probability distribution matrix, as shown in Table 8 .

4.2.2 Similarity Calculation

(1) Topic Probability Distance. Based on Table 8, we use Formula (3) to calculate the distance between two users' topic probability distributions, with results shown in Table 9 .

(2) Matrix Conversion. We convert this matrix to a similarity matrix using Formula (4), with results shown in Table 10 .

4.3 Recommendation Tag Generation

To intuitively display the tag recommendation process, this paper randomly selects user "enenn" and uses their annotation of book resource "The Chrysanthemum and the Sword" as an example. Based on the resource similarity (Table 6) and user similarity (Table 10) obtained above, we generate recommendations based on similar resources and similar users, and combine them with content-based tag recommendation to form the final tag recommendation result, thus explaining the entire recommendation tag generation process.

4.3.1 Recommendation Based on Similar Resources This process uses resource-based collaborative filtering. After calculating resource similarity and sorting it in descending order, we select the m1 most similar resources to target resource r R as its neighbor resource set, weight and sort the popular tags of these resources, and finally recommend the top n1 tags. Specific steps are:

(1) Select target resource r and sort its similarity with all resources in the system in descending order.
(2) Select the top m1 similar resources to generate neighbor resource set R'.
(3) Merge and weight-sort all popular tags in the neighbor resource set. For each tag t, its weight is shown in Formula (5):

W(t)= r R Sim(r,r')×Freq(t) 公式 (5)

In Formula (5), Sim(r,r') is the similarity between target resource r and resource r', and Freq(t) is the frequency of tag t. (4) For the sorted tags, select the top n1 as the recommendation result based on similar resources and normalize their weights.

Using the above steps, we obtain the target resource's neighbor resource set and candidate tag set A based on similar resources, as shown in Tables 11 and 12 .

By understanding the target resource "The Chrysanthemum and the Sword," we find it is a book describing Japanese culture and researching various aspects of the Japanese nation and character. Combined with the recommendation results in Table 12, we can see that tags recommended based on similar resources are closely related to the resource's content features, meeting users' basic needs during annotation. Since this method does not involve any user-related factors and is not limited by text content length, it has good stability. However, its recommendation results are relatively broad with high repetition, recommending only popular tags of resources in the system, with high socialization but no consideration of user interests, unable to recommend novel tags. Thus, recommendation results based on similar resources still have unavoidable defects.

4.3.2 Recommendation Based on Similar Users After calculating similarity between users, we find the m2 users most similar to target user u U who have annotated the target resource as the neighbor user set, then weight and sort the tags they used to annotate the target resource, and recommend the top n2 tags. Specific steps are:

(1) Select target user u and sort similarity with all users in the system in descending order.
(2) Select the top m2 users who have annotated target resource r as neighbor user set U'.
(3) Merge and weight-sort all tags used by neighbor users to annotate the target resource. For each tag t, its weight is shown in Formula (6):

W(t)= u U Sim(u,u')×Freq(t) 公式 (6)

In Formula (6), Sim(u,u') is the similarity between target user u and user u', and Freq(t) is the frequency of tag t. (4) For the sorted tags, select the top n2 as the recommendation result based on similar users and normalize their weights.

Using the above steps, we obtain the target user's neighbor user set and candidate tag set B based on similar users, as shown in Tables 13 and 14 .

Analyzing Table 14 and comparing it with Tables 12 and 15, this recommendation method considers user preference factors, thus recommending some personalized tags. For example, "United States" and "anthropology" do not appear in the recommendation results based on similar resources (Table 12), but considering that the author of "The Chrysanthemum and the Sword" is a famous American anthropologist and the book analyzes and studies the characteristics of the entire Japanese nation and Japanese personality, these two tags are very important for describing this book. Thus, tag recommendation based on similar users can improve tag novelty. Compared with content-based recommendation results (Table 15), this method has better result precision and more comprehensive content, but because it considers user factors, it has cold start and other problems.

4.3.3 Content-Based Recommendation Resource content feature information can intuitively reveal resource attributes and is an important source of recommended tags. Its advantage of not relying on user behavior information can well compensate for problems in collaborative filtering technology. This paper selects TF-IDF feature word extraction technology as the content-based tag recommendation method.

TF-IDF (Term Frequency-Inverse Document Frequency) is a weighted technique commonly used in information retrieval and text classification to evaluate the importance of a word to a document in a document collection. The basic idea of TF-IDF is that if a word appears very frequently in a document but rarely in other documents, the word is very important for that document and has good discrimination ability. The TF-IDF calculation formula for a word is shown in Formula (7):

Wij=tfij×idfi = (nij/ knkj) × log(N/ni) 公式 (7)

In Formula (7), nij represents the number of times feature word ti appears in document dj, and the denominator represents the total number of word occurrences in document dj. N represents the total number of documents, and ni represents the number of documents containing feature word ti. It can be seen that words with high TF-IDF values are usually the best terms for describing document content features. Specific steps are:

(1) Using the preprocessed book introduction content document collection containing various introduction feature words from Section 3.2, calculate the weight of all words using Formula (7).
(2) After sorting the calculation results in descending order, select the top n3 words as the content-based recommendation result and normalize their weights.

Through the above steps, we obtain content-based tag recommendation candidate set C, as shown in Table 15 .

Compared with the other two recommendation results (Tables 12 and 14), content-based tag recommendation can directly reveal various attributes of resources. For example, in Table 15, “Japan” and “culture” describe the resource’s content features, while “The Chrysanthemum and the Sword” and “Benedict” describe the resource’s external features of title and author. Therefore, its recommendation results are precise and not affected by user factors, effectively avoiding cold start problems. However, due to text length limitations and other reasons, its recommendation results are relatively narrow and not comprehensive enough, with some tags unsuitable for describing the book’s content, such as “contradiction.” Thus, this method may be more applicable for resources with sufficient text content.

4.4 Recommendation Result Generation After the above calculations, Tables 12, 14, and 15 respectively represent the three tag recommendation candidate sets based on similar resources, similar users, and content for target user

u and target resource r. These three results correspond to the three tag source angles in the tagging system: resource popular tags, user interest tags, and resource content tags. Since these three elements constitute the social tagging system, their importance is considered consistent. Therefore, when fusing these three recommendation results, we use a weighted calculation method, normalizing the weight coefficients of their respective results, then adding the weights and sorting in descending order, and selecting the top n4 tags as the final recommendation result to submit to target user u. Since Douban Reading recommends 10 tags to users, this paper also recommends 10 tags to users.

After calculation, for target user enenn annotating target resource "The Chrysanthemum and the Sword," the system's final recommended tag results are shown in Table 16 .

4.5 Result Evaluation and Analysis

To verify recommendation accuracy, we use Precision, Recall, and F1 value as evaluation indicators, as shown in Formulas (8)-(10):

Precision=TP/(TP+FP) 公式 (8)
Recall=TP/(TP+FN) 公式 (9)
F1=2×Precision×Recall/(Precision+Recall) 公式 (10)

In the formulas, TP represents the number of correctly recommended tags, FP represents the number of incorrectly recommended tags, and FN represents the number of tags that should have been recommended but were not.

We randomly selected 80% of the experimental data as the training set and the remaining 20% as the test set for prediction, and calculated the precision, recall, and F1 values of our experimental recommendation results. To further verify the effectiveness of the proposed method, we calculated and compared the values of several current tag recommendation methods (content-based recommendation, similar resource-based recommendation, similar user-based recommendation). For convenience, the proposed method is abbreviated as I+U+C, similar resource-based recommendation as Item-CF, similar user-based recommendation as User-CF, and content-based recommendation as Content-Based. The comparison results are shown in Figure 3 [Figure 3: see original paper].

Tables 12, 14, and 15 show the results obtained by the three single recommendation methods. Although the tags recommended by each method are highly consistent with the target resource "The Chrysanthemum and the Sword" and can clearly reveal the resource's characteristic tags, considering actual annotation situations, we can find that single recommendation methods all have unavoidable defects.

First, from the perspective of similar resource-based recommendation results (Table 12), this method starts from the perspective of popular tags of resources in the tagging system. Its recommendation results are stable and can basically describe and summarize resource features well. Since the source is popular tags, users have high acceptance, but precisely because of this, the results are too

socialized, with high tag repetition, lack of novelty, and limited user choice range.

Second, from the perspective of similar user-based recommendation (Table 14), the source of these results is the interest tags used by users in the system. It considers user preference factors, resulting in highly accurate, novel, and personalized tags. However, sometimes it overemphasizes user personality and is affected by it.

Third, from the perspective of content-based recommendation (Table 15), it starts from the resource content tag perspective, with tags directly extracted from the resource itself, not affected by socialization and user factors, with precise description closer to the resource itself. However, it is limited by various aspects of text content, and its recommended results are the fewest in number among several methods.

Compared with the three single methods, the experimental results of the method proposed in this paper are shown in Table 16, with relatively ideal recommendation results. This method mainly targets situations in social tagging systems where resources and tags have sufficient user information. When users prepare to annotate resources, it calculates relevant information from three perspectives: resource content, resource popular tags, and tags previously used by users, and recommends tags to them. However, it is not suitable for situations with particularly scattered information. Analyzing the results in Table 16, this method fuses the advantages of three recommendation technologies and corresponds to the three tag sources in the tagging system. The recommended tags reveal resource features from multiple perspectives: socialized tags from resource popularity (such as "Japan," "culture"), personalized tags from user interests (such as "United States," "anthropology"), and keywords from the resource itself (such as "Benedict"). The coverage is broad and relatively standardized, with good effects. At the same time, it avoids problems such as cold start and text limitations, expanding the range of tags users can choose from during annotation, allowing them to select appropriate tags based on their needs and understanding of the resource. From the comparison of experimental result data of various methods in Figure 3, we can also see that our method is superior to other methods in terms of precision, recall, and F1 value, proving that this method can effectively improve recommendation accuracy.

Douban is the most typical representative of social tagging websites in China. Whether domestic websites like Sina Weibo and Zhihu or foreign websites like Delicious and Flickr (images) all allow users to annotate resources with custom keywords like Douban. Therefore, although this paper only experiments on the Douban Reading dataset, as long as keywords are used as resource annotations, the method proposed in this paper is applicable and has certain generalizability.

Tag recommendation differs from general personalized recommendation as it needs to consider three factors: user, resource, and tag. Although existing recommendation methods use collaborative filtering ideas to varying degrees, they

simply use word co-occurrence as the basis for similarity calculation, ignoring semantic relationships among tags and affecting recommendation accuracy. This paper considers semantic information among tags and resource content, fuses three recommendation technologies corresponding to three tag source perspectives (resource content tags, resource popular tags, user interest tags), and uses LDA to calculate their respective distributions on topic probability from the topic semantic level as the data basis for similarity calculation, generating final recommendation results. The experimental results show that the hybrid tag recommendation method proposed in this paper improves similarity calculation accuracy through data dimensionality reduction and semantic relationship application, makes recommendation results socialized and novel, alleviates cold start and data sparsity problems, improves tag recommendation accuracy and tag quality, and achieves the purpose of tag standardization to a certain extent, providing a reference for future tag recommendation. However, to clearly explain the model, this paper uses a small amount of data in experiments, uses empirical values for some parameters, and does not consider differences in the number of different tag recommendations, which will inevitably affect the experimental results. Therefore, in subsequent research, we need to expand the data volume, test the effectiveness of this method on large-scale datasets, improve related algorithms, study the optimal parameter values for this model, optimize the recommendation process, improve recommendation result accuracy, and better move from theory to practice.

References

[1] Krestel R, Fankhauser P. Tag recommendation using probabilistic topic models[C/OL]//Proceedings of ECML PKDD discovery challenge (DC09), Bled, Slovenia, 2009: 131-141[2017-08-23]. https://www.kde.cs.uni-kassel.de/ws/dc09/papers/proceedings.pdf#page=131.

[2] Jin Yan, Chen Yu. Research on ontology-based tag control methods[J]. Library Theory and Practice, 2010(7): 26-29.

[3] Bogárdi-Mészöly Á, Rövid A, Ishikawa H, et al. Tag and topic recommendation systems[J]. Acta polytechnica hungarica, 2013, 10(10): 171-191.

[4] Fan Yongquan, Liu Yan, Lu Yuan. Review of research progress on social recommendation systems[J]. Modern Computer: Popular Edition, 2014(10): 29-33.

[5] Zhang Yin. Research on tag recommendation methods in social tagging systems[D]. Shenyang: Northeastern University, 2012.

[6] Qiao Lvyin, Zhang Min. Review of tag recommendation methods based on Folksonomy in China[J]. Journal of Information Resources Management, 2012(4): 41-46.

[7] Liu Zhili. Research on content-based social tag recommendation technology[D]. Harbin: Harbin Engineering University, 2012.

[8] Wang Guoxia, Liu Heping. Review of personalized recommendation systems[J]. Computer Engineering and Applications, 2012, 48(7): 66-76.

[9] Cai Y, Leung H, Li Q, et al. Typicality-based collaborative filtering recommendation[C]//IEEE international conference on tools with artificial intelligence, 2010, 2(3): 97-104.

[10] Tatu M, Srikanth M, Silva T. Tag recommendations using bookmark content[C/OL]//Proceedings of the ECML PKDD discovery challenge at 18th European conference on Machine Learning, Antwerp, Belgium, 2008: 96-107[2017-08-23]. https://www.researchgate.net/profile/Antal_{{{Van}}}{{Den}}{Bosch2}/publication/228075659_{{{Us
Language-Models-for-Spam-Detection-in-Social-Bookmarking.pdf#page=104.

[11] Mishne G. AutoTag: a collaborative approach to automated tag assignment for weblog posts[C]//Proceedings of the 15th international conference on World Wide Web. New York: ACM Press, 2006: 953-954.

[12] Marinho L, Schmidt-Thieme L. Collaborative tag recommendations[C]//Data Analysis, Machine Learning-Proceedings of the 31st Annual conference of the German classification society, Albert-Ludwigs-Universität Freiburg, Germany, 2008: 533-540[2017-08-23]. https://link.springer.com/chapter/10.1007%2F978-3-540-78246-9_{63}.

[13] Hotho A, Jäschke R, Schmitz C, et al. Information retrieval in folksonomies: search and ranking[J]. Lecture notes in computer science, 2006, 4011: 411-426.

[14] Song Hongxin. Research and implementation of blog retrieval experimental system based on tags and content[D]. Beijing: Beijing University of Posts and Telecommunications, 2011.

[15] Gao Bing. Research on tag recommendation technology in Q&A communities[D]. Harbin: Harbin Institute of Technology, 2009.

[16] Wang Chuanbao. Tag recommendation and search optimization based on collaborative filtering and text similarity[D]. Baoding: Hebei University, 2011.

[17] An Zhiwei. Research on tensor decomposition methods for social tag recommendation[D]. Changsha: Central South University, 2011.

[18] Zhang Liang. Research on tag recommendation method based on LDA topic model[J]. Modern Information, 2016, 36(2): 53-56.

[19] Li Huizong, Hu Xuegang, Yang Hengyu, et al. Comprehensive clustering method for social tags based on LDA[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(2): 146-155.

[20] Di Liang, Du Yongping. Application of LDA model in microblog user recommendation[J]. Computer Engineering, 2014, 40(5): 1-6.

[21] yhao2014. Popular understanding of LDA topic model[EB/OL].[2017-06-21]. http://blog.csdn.net/yhao2014/article/details/51098037.

[22] Zhang Peijing, Song Lei. Review of LDA-based topic modeling methods for microblog text[J]. Library and Information Service, 2012, 56(24): 120-126.

[23] Wang Zhenzhen, He Ming, Du Yongping. Text similarity calculation based on LDA topic model[J]. Computer Science, 2013, 40(12): 229-232.

[24] Zhong Qingyan, Su Yidan, Liang Shengyong. Research on tag recommendation based on hierarchical clustering and semantics[J]. Microcomputer Information, 2010, 26(36): 199-203.

[25] Wang Qian, Wang Junbo. An improved collaborative filtering recommendation algorithm[J]. Computer Science, 2010, 37(6): 226-228.

[26] Xiong Huixiang. Research on folksonomy for Web 3.0[D]. Wuhan: Central China Normal University, 2011.

[27] Douban Reading[EB/OL].[2017-06-05]. https://book.douban.com/.

[28] Shi Congying, Xu Chaojun, Yang Xiaojiang. Review of TFIDF algorithm research[J]. Computer Applications, 2009, 29(6): 167-170.

[29] Anand R, Jeffrey D. Big data · Internet large-scale data mining and distributed processing[M]. Beijing: Posts & Telecom Press, 2012: 6-7.

Author Contributions

Xiong Huixiang: Proposed the research direction, determined research methods and the logical framework of the paper.
Dou Yan: Collected data and wrote the paper.

Research on Tag Hybrid Recommendation Based on LDA Topic Model

Xiong Huixiang, Dou Yan
School of Information Management, Central China Normal University, Wuhan 430079

Keywords: social tagging; tag recommendation; collaborative filtering; LDA

Abstract: [Purpose/Significance] Current tag recommendation methods produce unsatisfactory results. This paper aims to improve traditional similarity calculation methods and combine various tag recommendation approaches to enhance recommendation accuracy. [Method/Process] Integrating content-based and collaborative filtering ideas, LDA is used to calculate similarity to find neighbor sets of resources and users, and keywords are extracted from resource content to construct a tag hybrid recommendation model. Finally, "Douban Reading" is used as an example to verify the model and compare it with several tag recommendation methods. [Result/Conclusion] In social tagging systems, three dimensions including user, resource, and tag should be considered. Only from a single angle will inevitably cause incomplete results. At the same time, the introduction of LDA in similarity calculation can exploit potential semantic

relationships and improve recommendation quality. And the combination of a variety of ways to learn from each other can make the results more satisfactory.

Note: Figure translations are in progress. See original paper for figures.