A Method for LDA-based Sina Weibo Recommendation

SangHao Xing
Zhejiang University City College
No. 51 Huzhou street
Hangzhou,ZheJiang
xshyzdyd123@gmail.com

Ziling Fan
Department of Biochemistry and Molecular and Cellular Biology, Georgetown University Medical Center
Washington, DC

ABSTRACT

Sina Weibo is one of the most influential social platforms in China. Recommendation system helps user to find celebrities that they may interest in and thus helps to attract more users. User's Weibo contents reflect their personal preferences. In this paper we proposed an LDA topic modeling based recommendation method which can discover topics of user's Weibo contents and recommend celebrities that users are interest in. The comparison result shows that our method outperforms tf-idf-based recommendation method.

CCS Concepts

• Computing methodologies →Data assimilation.

Keywords

LDA; SinaWeibo; Recommendation

1. INTRODUCTION

Sina Weibo is a Chinese microblogging service as twitter in western countries. In recent years, it has become one of the biggest and most popular social media platforms in China. Users can post, repost, and follow the users they are interested in. Because of its wide coverage of users, latest social topics and network properties, it also has become a valuable resource for various research fields such as topic analysis [8][9], information propagation analysis [10], network spreading analysis [2]. Weibo users follow celebrities, experts in different areas or any other weibo users who post weibo contents interest them. Therefore, to improve user experience, a precise and effective recommendation system which helps users to find potential followees based on personal preferences is an essential part of a complete application ecosystem.

Various Sina Weibo recommendation systems have been widely studied. Three recommendation systems are currently widely used which are content-based recommendation, user-based collaborative filtering and topology-Based recommendation system. Content-based recommendation [4] [7] have been extensively applied in existing applications. The algorithm uses tf-idf score to characterize users' Weibo contents and calculate users' preference by consine similarity. Term frequency (tf) is the frequency of a word in a document and inverse document frequency (idf) reflects the frequency of a word in a document set. This numerical statistic reflects how important a word to a document in a collection. Then the cosine similarity is calculated between documents (composed of contents of a potential followee) represented by a vector of tf-idf scores for words. The documents with high similarity with the target document is the potential recommendation. One primary shortcoming of this algorithm is that tf-idf only considers the lexical level feature between documents but cannot capture semantic features of them. Another commonly seen algorithm is called user-based collaborative filtering [1][5][11] which is a collaborative filtering algorithm based on the similarity between users. For example, let U represent a set of normal users selected randomly. A similarity score between target user and each of users in U will be calculated and recommendation will be made based on the followee list of those who have high similarity score with the target user. However, this algorithm is not applicable to a new user due to the cold start issue. Topology-Based recommendation system [6] is another widely used recommendation system. It explores the connections starting at the target user in order to select a set of candidate recommendations. It can be accomplished with the following steps: 1.Get the set of the target user's followees and denote this set S. 2.Get the followers of each element in S and denote this set L. 3.Get the followees of each element in L and denote this set T and the candidate recommendations will be subtraction of T from S. However, the assumption of this algorithm is that the target user is similar with the followers of his followers. Therefore, it may recommend user that the target user is not interested in.

We considered applying LDA model to our recommendation. Latent Dirichlet Allocation (LDA) is an unsupervised topic model. We assume that users' preference is reflected by their posts and reposts. Sina Weibo recommendation needs to find users' hidden preferences and LDA model can complete this task. We picked celebrities that Sina Weibo provides and used their Weibo contents to train LDA model. Then, we used the already trained LDA model to extract topics that common users interested in and recommend celebrities in these topics. Celebrities' Weibo contents. As a comparison, we used tf-idf-based recommendation to compare with our recommendation method. Common users' Weibo contents can be charactered by tf-idf weight. We compared tf-idf-based recommendation and our recommendation. The result shows that our recommendation can get hidden preferences more accurately.

This rest of this paper is organized as follow. In Section 2, we describe the method of data collection and data preprocessing. We introduce our recommendation and tf-ifd -based recommendation and propose method to compare our recommendation and tf-idf-based recommendation. In Section 3, we discuss the result of our experiment.

BDET2020, January 3–5, 2020, Singapore.
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7683-9/20/01…$15.00
DOI: https://doi.org/10.1145/3378904.3378913

2. METHODS

2.1 Weibo Collection and Data Preprocessing

We selected two types of Sina Weibo users: celebrities and common users. Celebrities are influential users who have a large number of followers. Common users should meet two requirements: the number of common user's followee is limited to range of between 300 and 600 and the number of post must be greater than 1000. Both of them should be active users who posted at least within a week.

We selected celebrities from ranking list that Sina Weibo provided. This ranking list is calculated according to the number of post, re-post and new followers within a month. It includes different types of topics, such as entertainment, sports, military and we selected 14 topics. For each topic, we picked 25 the most influential celebrities from this ranking list.

We extracted contents of users using the Sina Weibo API. The contents include all post and re-post of a Weibo user.

We did data preprocessing in this part, we tokenized word by using Jieba that a Chinese word tokenization module. We removed stop words and single word to achieve a clean collection of words or phrases from user weibo content.

2.2 Latent Dirichlet Allocation (LDA) for Weibo Recommendation System

LDA (Latent Dirichlet allocation) is proposed by [3] and commonly used as a topic modeling method. The model can give the probability distribution of the topic of each document in our document set. Meanwhile, it's a typical bag of words model. The bag of words model treats a document composed of a group of words and there is no ordinal relation between words and words. LDA is a three-layer Bayesian model: the core of LDA model is the distribution of topics obeys a Dirichlet distribution with a parameter α and the distribution of words obeys Dirichlet distribution with a parameter β . For each document, the generation process of the *i*th word of the document is as follows:

Choose topic Z that according to distribution θ.

Use the topic Z to generate the word according to distribution φ .

LDA model can be described as a probabilistic graphical:

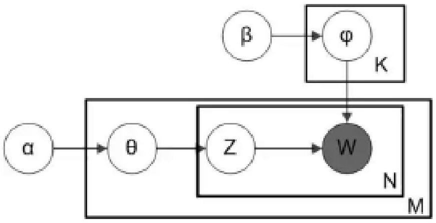

Figure 1. Probabilistic Graphical of LDA Model.

Our goal is to recommend some celebrities to common users and we applied LAD to our recommendation algorithm. The LDA based recommendation algorithm can be described as following steps:

1.We chose 350 celebrities as potential recommendation. these celebrities are divided into 14 groups, each group represent a topic category mention above and 25 celebrities per group. 50 common users that meet criteria mentioned above.

2. We can categorize all of words in the whole set of documents into a certein number of categories provided as a parameter for the algorithm and rank the words in each category based on the possibility of it in the category. Then we can annotate each category according to the words of each category.

3.we used the trained LDA to calculate the proportion of words assigned to each topic within a user Weibo content for each common user. The trained LDA model can calculate the distribution of topics in a new document and We selected the 3 highest ratio topics to the common user as recommendation topics. For each topic in recommendation topics, we chose 3 celebrities to recommend to common user.

2.3 TF-IDF Recommendation Algorithm

We selected tf-idf based recommendation as a method to compare with and evaluate our method. Tf-idf is the product of the tf and idf. TF is the frequency of a word in a document. TF can be computed the following:

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF reflects the frequency of a word in a document set .IDF is defined as the following:

IDF(t) = log (Total number of documents / Number of documents with term t in it).

Then we can get the tf-idf vector representation of each document in the document set. We calculated the similarity between celebrities and each common user. For each common user, we selected top 10 celebrities with the highest similarity as a recommendation list.

2.4 Evaluation Method

We evaluated our method using following two metrics: 1) overlap rate with term frequency ranking and 2) comparison with tf-idf recommendation method.

First, we compared the overlap rate of same terms ranked by LDA and term frequency count in each category. We selected top 30, top 50 and top100 terms ranked by our fitted LDA model. Then we also extracted the 30, 50 and 100 highest frequency words for each topic by using term frequency count. For each topic, we calculated the overlap of term set that LDA model generates and the term frequency.

Second, we evaluated our method by comparing it with tf-idf recommendation method which is a common content-based recommendation system. We compared them from two perspectives. For each common user in our user set, we extracted all celebrities from common user's followee list and sorted these celebrities in a descending order by the number of followers of these celebrities. We seeked the ground truth of user interest by selecting top 10 celebrities(ranked by follower number) from each user followee list and annotate the interest categories by counting the number of celebrities in each category. We predicted user interest using fitted LDA model as described above. Then we calculated overlap percentage between topics that LDA predicted and the ground truth. For tf-idf-based recommendation, we selected the 10 celebrities with the highest similarity as a recommendation list and annotate topic categories of these celebrities manually based on their post content. In this prepective, for each common user, we used the ground truth as described above perspective, we used LDA model to predict user interest and calculated the percentage of topics recommended by LDA in the ground truth. For tf-idf-based recommendation, we annotated topic categories of 10 celebrities that with the highest similarity and calculated the percentage of topics recommended by similarity of tf-idf weight in the in the ground truth.

3. RESULTS

3.1 Overlap of Selected Terms from Each Topic between LDA Method and Term Frequency Analysis

In this section, we discuss our experimental result. LDA model can generate important terms in each category. We compared LDA and term frequency count by calculating the overlap rate of the same terms using LDA and term frequency count. In table 1, the first column is 14 topics that we picked from the ranking list that Sina Weibo provides. But in our LDA models, there are 3 topics that we cannot annotate from the LDA model result and we removed these topics. For each topic, we calculated separately the overlap of top-30 terms, top-50 terms and top-100 terms between LDA and term frequency count. For example, we show one topic category in table 2. We selected top 30 terms and the same terms expressed in bold. These two methods have numerous repeated terms. there are several terms generated by LDA are related to the topic, but the term set that term frequency generates does not have these terms. Like "运动"(sport), "健身"(fitness), and "减肥"(lose weight)". Therefore, compared to use terms frequency count, the terms generated by LDA are more relevant to topic. Table 1 shows the term set that calculated by LDA and ranked by possibility. Top-30, top-50 and top-50 term set has plenty of high frequency terms. The last row shows that as the number of topic increases, the average of overlap rate is performance increasingly.

Table 1. The Percentage of Overlapped Terms Selected by LDA and Term Frequency

Topic category	Overlap of top-30	Overlap of top-50	Overlap of top-100
Military	73.3%	84%	76%
Electronic games	83%	74%	82%
Health	70%	70%	72%
Pet	70%	82%	70%
Emotional	76%	84%	88%
Finance	77%	84%	82%
Sport	70%	80%	75%
Beauty products	80%	78%	84%
Travel	73.3%	80%	77%
Entertainment	73.3%	76%	77%
History	73.3%	76%	74%
Average	**74.5%**	**78.4%**	**79.5%**

Table 2. An Example of The Result of Comparing LDA with term frequency

Category	Terms generated by LDA	Terms generated by term frequency	Percentage of overlap
Health	**健康** 可以 运动 **视频** **养生** 动作 **微博** 健身 **自己** 训练 **食物** **身体** 减肥 **一个** **如果** **每天** 分钟 不要 **这个** **饮食** **一定** **起来** **营养** 需要 简单 **容易** 每个 **没有** **就是** **生活**	**健康** **养生** **可以** **视频** **微博** **食物** **自己** **一个** 中医 我们 **身体** **营养** **没有** **如果** **就是** **生活** **每天** **这个** 旅游 孩子 什么 **起来** 维生素 可能 **一定** **容易** 很多 因为 **饮食** **需要**	70%

Table 3. An Example of User Followee Category Annotation

Top 10 celebrities	Topic categorize of celebrities	Main topic categories
1 财经网 2 新浪财经 3 新浪视频 4 天鹅娱乐 5 王芊灵 6 一手 Video 7 影视爆料者 8 雯蕾 9 捡书大叔, 10 搞笑柏柏	Finance Finance Entertainment Entertainment Entertainment Entertainment Entertainment Emotional Emotional Humor	Entertainment Finance Emotional

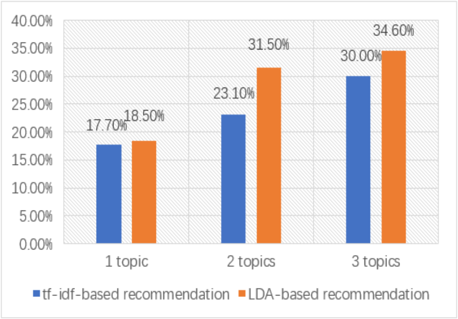

Figure 2. The Percentage of Overlapped Terms Selected by LDA and Term Frequency.

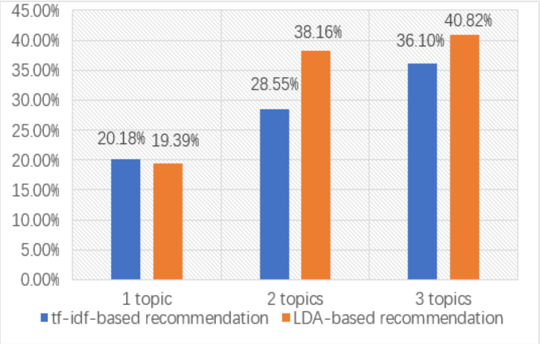

Figure 3. Proportion of The Topics Recommended by The Two Recommendation Systems in The Topics of Celebrities That Users Follow.

3.2 Comparison of LDA Based and Tf-Idf Recommendation Methods

In this part, we compared LDA-based recommendation and tf-idf-based recommendation. For each common user, we selected 10 celebrities from their followee list and annotated topic categories of these celebrities as described above. These categories are users truly interested in. We show one of the results of this step in table 3. First column is 10 celebrities, second column is Topic category of celebrities annotated by us and the last column is 3 main topic categories of these celebrities calculated by counting the number of celebrities in each category and selecting the top three categories. First, we compared different number of topics recommended by LDA-based recommendation and tf-idf- based recommendation. There are 1 topic, 2 topics, 3 topics. We calculated the percentage of overlap between the topics recommended by two recommendation systems and the topics that users truly interest in. Figure 2 shows that the result of this method. Regardless of the number of recommended topics selected, LDA-based recommendation is more accurate than tf-idf-based recommendation. As the number of recommended topic increases, the percentage of overlap also increases.

Similar to the previous method, we calculated proportion of the topics recommended by the two recommendations from the topics of celebrities that users follow. Figure 3 shows that LDA-based recommendation is more accurate than tf-IDF-based recommendation in general. When we select 1-topic, the accuracy of tf-idf-based recommendation is 20.18% and the accuracy of LDA-based recommendation is 19.39%, tf-idf- based recommendation performances higher accuracy rate. Both of two recommendations are more accurate with the number of recommended topic increases.

4. CONCLUSION

In this paper we proposed a LDA based sina weibo recommendation system. The method utilizes the content of official categorized weibo celebrity accounts to create LDA topic models. The categorized topics are annotated manually based on representative terms selected by the method and then the model is used to predict users interests based on the contents (posts or reposts) of users account. Finally, the recommendation is made based on the predicted categories given by the model. Experiments on Sina Weibo data for our recommendation and tf-idf-based recommendation showed that our recommendation gave more accurate recommendation compared to tf-idf-based recommendation.

5. ACKNOWLEDGMENTS

I would like to extend my sincere gratitude to my supervisor, Ziling Fan, for his instruct advice and useful suggestions on my thesis. I am deeply grateful of his help in the completion of this thesis.

6. REFERENCES

[1] B. Wang, J. Huang, L. Ou, and R. Wang. A collaborative filtering algorithm fusing user-based, item-based and social networks. In *IEEE Bigdata*, 2015, pp. 2337–2343.

[2] C. Liu, X. X. Zhan, Z. K. Zhang, G. Q. Sun, P. M. Hui. "How events determine spreading patterns: information transmission via internal and external influences on social networks." *New Journal of Physics* 17.11 (2015): 113045.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] I. Cantador, A. Bellog ń, and D. Vallet. Content-based recommendation in social tagging systems. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 237-240. ACM, 2010.

[5] I. Konstas, V. Stathopoulos, and J. M. Jose. On social networks and collaborative recommendation. In *SIGIR*, pages 195–202, 2009.

[6] Marcelo G. Armentano, Daniela Godoy, and Analia Amandi. Towards a followee recommender system for information seeking users in twitter. In *Proceedings of the Workshop on Semantic Adaptive Social Web*, pp. 27–38, 2011.

[7] M. Balabanovic and Y. Shoham. Fab: Content-based, collaborative recommendation. *Communications of the Association for Computing Machinery*, 40(3):66–72, 1997.

[8] N. Zheng, S. Song, and H. Bao. A temporal-topic model for friend recommendations in Chinese microblogging systems. *IEEE transactions on systems, man, and cybernetics: systems*, vol. 45, no. 9, pp. 1245– 1253, 2015.

[9] S. Wang, M. J. Paul, and M. Dredze. Exploring health topics in Chinese social media: An analysis of Sina Weibo. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*. pp. 20– 23, 2014.

[10] Wu M, Guo J, Zhang C, Xie J. Social media communication model research bases on Sina-weibo. Knowledge Engineering and Management, 2012: 445-454.

[11] Z. Zhao and M. Shang. User-based collaborative-filtering recommendation algorithms on hadoop. In *KDD*, pages 478–481. ACM, 2010.