A Recommendation Model Based on Content and Social Network

Hang Xue
School of Computer Science
Beijing University of Posts and Telecommunications
Beijing, China

Dongmei Zhang
School of Computer Science
Beijing University of Posts and Telecommunications
Beijing, China

***Abstract*—Recommendation model are a popular trend in recent research in Internet technologies. However traditional content-based recommendation, social network-based recommendation and collaborative filtering recommendation have their own shortcomings. To overcome them, we proposed a recommendation model based on content and social network (RMBCS). First, we proposed a new distance to calculate the text similarity between long text and short text. Then, we proposed a new method to find the nearest neighbor group from user's social network quickly and conveniently. At last, we recommend the texts which user's nearest neighbor group had read to the user. Experimental results indicate that our model has a better performance.**

***Keywords*—recommendation model; content-based recommendation; social network-based recommendation; text similarity; nearest neighbor group**

I. Introduction

With the maturing of Internet technology, people have entered the era of big data, it is difficult for users to quickly find their own information from a large amount of data information, different users use search engines often get the same recommended results, but users would prefer to be personalized recommendations according to their interests and preferences. Many scholars have proposed a variety of recommendation algorithms, among which collaborative filtering (CF) algorithm is widely used in recommendation system because of its good extensibility and realizability. However, there are many drawbacks in the recommendation of collaborative filtering, which mainly include: (1) Data sparse problem. Due to the sparsity of the user-item scoring matrix data, the calculated user or project similarity must be inaccurate, which in turn affects the recommendation accuracy and the user experience. (2) Cold start problem. In the recommendation system, since the new user does not have the rating information for the project, and the new project is not scored by the user, it is impossible to calculate the corresponding nearest neighbor, so it cannot be recommended. (3) The calculation of the traditional collaborative filtering algorithm can only distinguish the degree of interest between users, but it cannot distinguish between the friend relationship and the strangeness between users. (4) Finding the cost of the nearest neighbor user group is too high, both the time cost and the running cost of the machine are large. When the number of users is massive, the cost is difficult to estimate[1], [2].

The content-based recommendation (CB) model is a model that is specifically recommended for unstructured data such as texts and videos. It is highly interpretable, and the recommended results are more easily trusted by users; even if the items are not evaluated by users (score, browse, collect etc.) New items can also be recommended with a higher accuracy rate. However, it also has the following disadvantages, mainly including: (1) The recommended content is overly consistent. The content-based recommendation system is recommended according to the user's evaluation history, so that the recommended items may be too similar to the user's rated items, resulting in lack of novelty; (2) Cold start problem is serious. When the new user does not have any operation records, the system can not recommend new item to the new user[3].

On the other hand, many recommendation systems are based on the assumption that the user's evaluation or behavior of the project is independent of each other, and users subjectively score, collect, or like the project according to their interests, without being influenced by other users. In real life, however, most users generally ask for advice from friends they trust or close friends on social networks when they decide to choose a project, and people's hobbies are often influenced by friends in the circle of friends or loved ones in reality. The social relationships of users partly reflect how similar they are to their hobbies and how similar they are to the real world[4].

The social network-based recommendation just makes up for the above shortcomings, but the social network-based recommendation system also has the following disadvantages: the offline accuracy of the recommendation algorithm is not necessarily improved, because the friend relationship in the user's social network is not generated based on the common interest user. Therefore, the interests of user friends are often inconsistent with the interests of users. So how to find a user group that is closest to the user is the top priority of recommendation[5].

978-1-5386-8178-7/19/$31.00 ©2019 IEEE

How to reasonably combine the content-based recommendation algorithm and the recommendation algorithm based on social network, and use the advantages of both parties to complement their shortcomings to generate personalized recommendation for text, which is of great significance to the improvement of recommendation system accuracy and user experience.

II. The Recommendation Model Based on Content and Social Network

A. Feature extraction

The extraction process of the features used in this paper is shown in figure 1. In the text data set, the text is first classified according to the length of the text, and is divided into short texts containing 0-100 words and long texts containing 100 words or more.

Since the Chinese text is processed, the word segmentation process is first required. At the same time, the latest dictionary is used in word segmentation, including network terms, new words and so on, in order to ensure the correctness of word segmentation and the accuracy of text feature extraction.

Because the short text length after the word segmentation is small, the amount of information of each word is larger than that of the long text word, so only the stop words are deleted for the short text. And because there are so many rich meanings in the short text, we do the process of restoring the expression, and manually remark the meaning of the expression with text to keep the short text information as much as possible. Stop words and single words are removed in the process of filtering, and the result of the word segmentation is finally obtained.

The LDA topic model is used to extract the topic words from the word segmentation results of the text, and the keyword is transformed into the word vector using word2vec for later calculation. For the user's characteristics, the user's history and behavior, such as all the texts that the user has bookmarked and liked, are extracted as the characteristics of the user.

B. Measurement of similarity

1) Similarity between short text and short text: Short text is for social platforms such as chat software, forums, messages, text messages. Because the length of short text is from dozens of words to about 100 words, its characteristics are not obvious.

The effect of topic models on short texts is less than ideal, and the application of word vectors on short text-short text matching tasks is more common than topic models. So here use Word2vec to convert words into word vectors and use cosine distance (CD) to calculate the distance.

$$S_{ss}(\mathbf{P},\mathbf{Q}) = \frac{\mathbf{P}\cdot\mathbf{Q}}{\|\mathbf{P}\|\,\|\mathbf{Q}\|} \tag{1}$$

where $\mathbf{P}$ and $\mathbf{Q}$ represent vectors of two short texts, respectively.

2) Similarity between long text and long text: By using LDA, we can get a thematic distribution of two long texts, and then measure the similarity between them by calculating the distance of two multiple distributions. Here we use Heilinger distance (HD) to measure the distance of two multiple distributions.

$$S_{ll}(\mathbf{P},\mathbf{Q}) = \frac{1}{\sqrt{2}}\left\|\sqrt{\mathbf{P}}-\sqrt{\mathbf{Q}}\right\|_2 \tag{2}$$

where $\mathbf{P}$ and $\mathbf{Q}$ represent the distribution of the themes of the two long texts, respectively.

3) Similarity between long text and short text: This paper mainly proposes a method for measuring the similarity between long text and short text which is named generating distance (GD). GD can not only be applied in the recommendation system of this paper, but also plays an important role in real life. For example, generally the query in the search engine is short text, but the results of the query (webpages) are mainly long texts. The long-short text similarity calculation can give excellent query results even instead of using the search matching algorithm.

The similarity between short text and long text is different from short-short text and long-long text mentioned above. When calculating similarity, we avoid to direct mapping of short texts, but the topic distribution based on long texts. We calculate the probability that the distribution generates short text as the similarity of long-short text, which is called generating probability (GP):

$$S_{ls}(\mathbf{P},\mathbf{Q}) = \left(\prod_{r\in\mathbf{P}}\sum_{k=1}^{K} p(r|z_k)p(z_k|\mathbf{Q})\right)^{\frac{1}{cnt(\mathbf{P}))}} \tag{3}$$

where $\mathbf{P}$ represents the feature vectors of the short text, $\mathbf{Q}$ represents the feature vectors of the long text, r represents the word after the word segmentation in the short text, and z_k represents the kth topic, $cnt(p)$ represents the number of vectors contained in $\mathbf{P}$.

However, since $p(r|z_k)$ is often affected by high-frequency words, some semantically important low-frequency words are difficult to be selected as feature words, cosine distance calculation for r and z_k will have a better effect. This is how generating probability converts to generating distance.

$$S_{ls}(\mathbf{P},\mathbf{Q}) = \left(\prod_{r\in\mathbf{P}}\sum_{k=1}^{K} cos(\mathbf{r},\mathbf{z_k})p(z_k|\mathbf{Q})\right)^{\frac{1}{cnt(\mathbf{P}))}} \tag{4}$$

where $\mathbf{r}$ represents the word vector corresponding to r , and $\mathbf{z_k}$ represents the word vector corresponding to z_k.

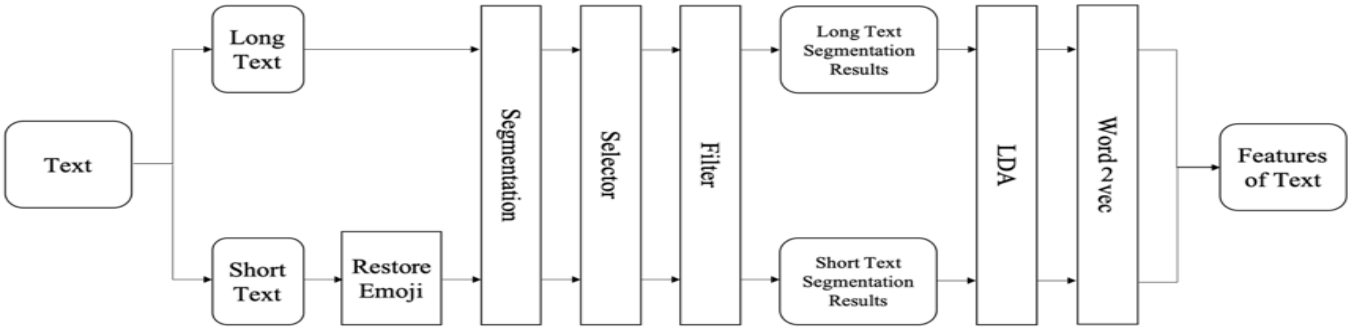

Fig. 1. Process of Texts Segmentation Results

C. The choice of the nearest neighbor user group in the social network

First, we introduce the concept of isomorphic group (IG), heterogeneous group (HG) and outland domain (OD).

IG represents a direct friend of the user in the system, and a cluster directly connected to the central user in the social relationship, which represents a set of points in the social network that have a distance of 1 from the center point. HG refers to a cluster that is only indirectly connected through a homogeneous group, which represents a set of points in the social network that are 2 points away from the center point. OD represents a collection of points in the social network that have a distance of 3 from the center point.

After establishing a social network, we traverse the social network centered on each user, find their IG, HG and OD. By comparing them using user's similarity, we find the nearest neighbor as the user group that recommends text to the user. Through experiments we found that it is the most efficient to find the nearest neighbor user group from HG, and the accuracy rate is close to the nearest neighbor user group found by the collaborative filtering.

III. Experimental results and analysis

A. Data Set Description

Since there is no suitable data set, we crawled the data from the Internet (Zhihu and Weibo). The data set is divided into three parts:

The first part is the user data, including basic attributes such as the number of favorites, the number of attentions, etc. There is a total of 157,377 data; the second part is the user's social network data, including the attention and attention of the users, a total of 122,166 Data; the third part is text data, which contains a total of 12,899,495 data.

In the user data set, we cleaned the data where user's collection number is less than 10, or the number of attention and the total number of attentions is less than 20, and the attribute information of the user is almost empty. After that 12,2166 pieces of data are obtained. The text data set removes the link information of the picture or text, filters the text with low length's text and some empty data, and gets 10,248,342 data.

B. Evaluation method

- Accuracy

$$A = \frac{1}{N}\sum_{i=1}^{N}\frac{G(Y_i)\bigcap F(\hat{Y_i}))}{G(Y_i)} \tag{5}$$

N is the number of users, $F(\hat{Y_i})$ represents all the text recommended by the system for user i prediction, $G(Y_i)$ represents the true historical behavior data of user i.

- Similarity

Through the previous text similarity, we know that the similarity of text is calculated by using different calculation methods for different categories, and the maximum value after traversing the recommended text is the similarity of the text of the current real collection. So, the average similarity of the system as a whole is:

$$S = \frac{1}{M}\sum_{i=1,j\in J}^{N} max(s(\mathbf{Y_i}, \hat{\mathbf{Y}}_\mathbf{j})) \tag{6}$$

- Novelty

$$N = \frac{C(\hat{Y}\bigcap Y)}{C(Y)} \tag{7}$$

$C(\hat{Y})$ represents the number of subject classes for all text in a real collection, and $C(\hat{Y}\bigcap Y)$ represents the number of subject classes for the correct text in the recommendation.

C. Comparison method

As mentioned above, in the calculation of text similarity, we use cosine distance, Heilinger distance and Jaccard distance to compare; in the recommended effect of the model, content

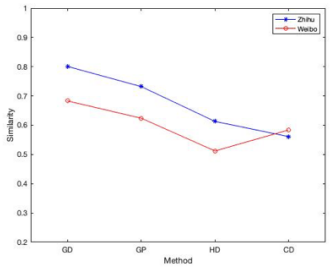

Fig. 2. Average Similarity

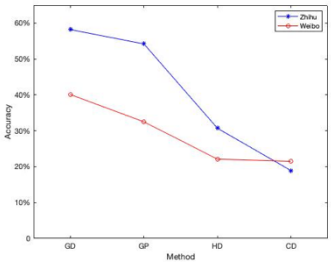

Fig. 3. Average Accuracy

and social network based (CSB) recommendation model is compared with CB and CF.

D. Experimental results

1) Accuracy of GD: It can be seen from figure 2 and figure 3 that the similarity and accuracy of the generating distance are obviously improved in the marked data set compared with other methods mentioned above. It is enough to illustrate that the proposed generating distance has certain effects on calculating similarity between long text and short text.

2) Selection of the nearest neighbor group: We have calculated the similarity of users from HG, IG, and OD for the Zhihu and Weibo data sets and compare them from the average similarity of all users, the top 5 and top 10 in the group. As can be seen from figures 4 and 5, the similarity of IG is much higher than that of HG's, regardless of the dataset and the average value. Although the accuracy of IG is slightly lower than that of OD, it can be seen from figure 6 that the training time is much lower than the training time of OD. So, in order to ensure accuracy and low overhead, we believe that choosing the nearest neighbor user group from IG is the best.

3) Accuracy of recommendation: We calculated the recommended accuracy in the Zhihu and Weibo data set. Obviously, as we can see from the figure 7 and 8, as the number of recommended texts increases, the accuracy rate is rising. The accuracy of CSB is much better than CB, and slightly lower than CF.

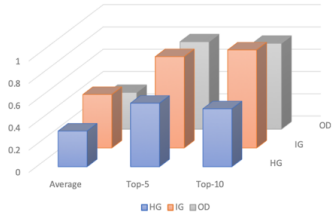

Fig. 4. Similarity of Zhihu

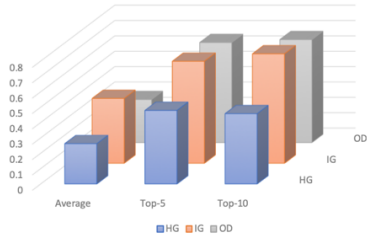

Fig. 5. Similarity of Weibo

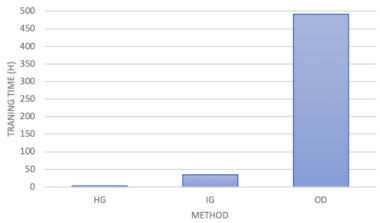

Fig. 6. Training Time

4) Similarity: From the perspective of text similarity, the similarity of CSB is the highest among the three methods in the figure 9, which also shows that the calculation method of text similarity proposed in this paper is better, and it can recommend texts which are more similar with users' history records.

5) Novelty: In terms of novelty, CSB has the best novelty on both data sets in the figure 10. That is to say, CSB can not only achieve excellent results in accuracy and similarity, but also recommend various types of texts. This is also more in line with user needs.

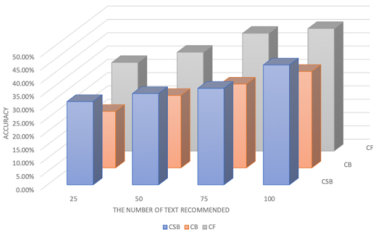

Fig. 7. Recommendation Accuracy of Zhihu

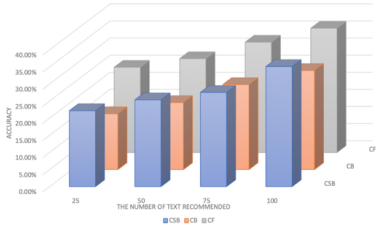

Fig. 8. Recommendation Accuracy of Weibo

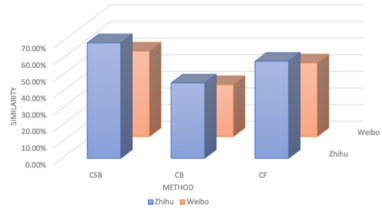

Fig. 9. Recommendation Similarity
(The number of recommended texts = 50)

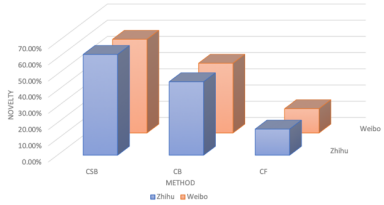

Fig. 10. Recommendation Novelty
(The number of recommended texts = 50)

IV. Conclusion

In this paper, we propose a recommendation model based on content and social network. The model uses the LDA to extract user and text features and use word2vec to turn words into word vectors. Using text similarity comparison for the features of each text, make the text with high text similarity as the same class. This paper also proposes a method called generating distance to calculate the similarity between long text and short text. Through experimental verification, we find generating distance is better than the existing method of calculating the similarity of texts, and has a good matching. It also completes the vacancy of the entire text similarity calculation. The RMBSC compares the user's IG, HG, and OD in the social network, and finds the user's nearest neighbor group which is used to recommend is in the HG. By recommending unread articles from the user's nearest neighbor group, this not only improves the accuracy of the recommendation, reduces the cost and cost of training, but also enhances the novelty of the recommendation and more in line with the needs of the user. At the same time, different methods are used to deal with different situations in the cold start problem to slow down the cold start problem.

However, due to the limitation of data crawled, we can't get the data of the real new users, so it is impossible to verify from the experimental point of view whether it can really reduce the effectiveness of the cold start problem.

The future work is to optimize the similarity of text, for example, the calculation of text similarity can be compared with other calculation methods, to find a more effective matching method to further improve the recommended performance.

Acknowledgment

The authors would like to thank the anonymous reviewers whose insightful comments have helped improve the presentation of this paper significantly.

References

[1] X. Yang, Y. Guo, Y. Liu, and H. Steck, "A survey of collaborative filtering based social recommender systems," *Computer Communications*, vol. 41, pp. 1–10, 2014.

[2] G. Xu, Z. Wu, Y. Zhang, and J. Cao, "Social networking meets recommender systems: survey," *International Journal of Social Network Mining*, vol. 2, no. 1, pp. 64–100, 2015.

[3] J. Su, "Content based recommendation system," Jan. 5 2016, uS Patent 9,230,212.

[4] J. H. Errico, M. I. Sezan, G. R. Borden, G. A. Feather, and M. G. Grover, "Collaborative recommendation system," Feb. 3 2015, uS Patent 8,949,899.

[5] S. Amer-Yahia, A. Galland, R. Yerneni, and C. Yu, "Recommendation system using social behavior analysis and vocabulary taxonomies," Nov. 24 2015, uS Patent 9,195,752.

[6] S. Sedhain, S. Sanner, D. Braziunas, L. Xie, and J. Christensen, "Social collaborative filtering for cold-start recommendations," in *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 2014, pp. 345–348.

[7] T.-H. Lin, C. Gao, and Y. Li, "Recommender systems with characterized social regularization," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 2018, pp. 1767–1770.

[8] D. Cohen, M. Aharon, Y. Koren, O. Somekh, and R. Nissim, "Expediting exploration by attribute-to-feature mapping for cold-start recommendations," in *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 2017, pp. 184–192.

[9] W. Li, M. Gao, W. Rong, J. Wen, Q. Xiong, R. Jia, and T. Dou, "Social recommendation using euclidean embedding," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 589–595.

[10] X. He, H. Zhang, M.-Y. Kan, and T.-S. Chua, "Fast matrix factorization for online recommendation with implicit feedback," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2016, pp. 549–558.

[11] C.-K. Hsieh, L. Yang, Y. Cui, T.-Y. Lin, S. Belongie, and D. Estrin, "Collaborative metric learning," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 193–201.