Review of Friend Recommendation Algorithms Based on User Text Data

Yuxin Qu

Abstract— **Social networks are a powerful complement to the interaction needs of human society. With the development of the Internet, people turn offline dating into online dating and expand our circle of friends. The rapid development of the Internet has brought us a lot of convenience, but also caused the problem of information overload, recommendation algorithm, filter some unnecessary information, only to provide us the information we need. The main work of this paper is to study how to obtain effective information from a large amount of text data and use this information to recommend friends to target users. In this paper, Latent Dirichlet Allocation topic model is used to extract the topic from text data, in order to characterize the behavioral characteristics of target users, and then make friend recommendation for them.**

Index Terms—**Social network, recommendation algorithm, LDA theme model, friend recommendation.**

I. Introduction

Since its birth, the Internet has been built for the convenience of human social contact. Socializing is a need for all of us. While our traditional social life is often limited by space and time, you never know that there will be someone like you somewhere in the world, but due to the limitations of time and space, it is difficult to find this person. The birth of the Internet has broken this limitation, and people can socialize with people anywhere, anytime.

In 1971, in order to facilitate the exchange of research results, the scientists of ARPANET project created the first E-mail, and people began to use E-mail to make friends[1]. People moved their social relationships in real life to the Internet, and it was easier to make new friends and expand their circle of friends through the Internet. With the development of Internet technology, the Internet has rapidly developed into the era of Web2.0, which is an era of interconnection[2]. In terms of interpersonal communication, social network has greatly expanded the scope of people's communication, and a large number of social software has emerged. Find new friends on Tianya community forum, QQ, Renren, Jiayuan, Douban, Sina Weibo, Facebook, Twitter and so on. More and more people use the Internet to make friends, but a large amount of data information is generated. While enjoying the convenience brought by the Internet, people need a lot of time to obtain valuable information for themselves, and also suffer from the trouble of information overload[3].

The recommendation system is produced to solve the problem of information overload. The recommendation system is a kind of information filtering system. For example, if the user wants to watch a movie but is not sure which one it is, the algorithm will recommend the user according to the user's previous preferences. A lot of text data are generated in our daily online surfing, including chat files between us and other users, etc. How to recommend friends for users through these data is the main task of this paper.

Manuscript received July 21, 2022.

Yuxin Qu, The school of Software, Tiangong University, Tianjin, CHINA.

II. Related Work

Content-based Recommendations (CB) focuses on the metadata of the project and makes use of the characteristics of the project to find and recommend items similar to those once loved by target users[4]. It uses a machine learning algorithm to induce a profile of the users preferences from training examples based on a feature description of content. Based on the user profile learned, items that are matched to the profile will be recommended to the active user. For example, users express interest in the movie Source Code, We'd recommend the RESET. Usually applied to description by the area of the project, the basic idea is by analyzing the user's attribute, and text information, unearthed user interests related keywords or tags, and then use this knowledge to build user interest profile and interest in these files and product feature matching and make recommendations. As shown in Figure 1, content-based recommendation algorithms can generally be divided into the following three steps:

Figure 1 Content-based recommendation algorithm flow

Collaborative Filtering Recommendations (CF) is one of the mature and common methods. It does not need to use the user's or item's archives, but only needs to collect the user's historical behavior records[5]. There is a database consisting of users and their ratings for the variety of items which have been seen by them. The user who wants recommendations is called active user. If we want to recommend items for an active user, it simply searches the database to find other similar users for the active user. Based on those similar users, it will recommend the items. Explore the potential similarity between users and items, and make recommendations based on the similarity of this group. As we have learned above, content-based recommendation algorithms predict the possible preferences of the future based on the previous data

of the target user. On the contrary, collaborative filtering recommendation algorithms are based on the assumption that users with consistent opinions in the previous selection will also make the same choice later. As shown in Figure 2, the process of collaborative filtering recommendation algorithm can be divided into the following three steps:

Figure 2 Collaborative filtering recommendation algorithm flow

The hybrid recommendation algorithm integrates the advantages of each individual recommendation algorithm, and can also solve the problems existing in the single recommendation algorithm, and improve the recommendation accuracy and performance of the recommendation system. The hybrid recommendation algorithm combines two or more recommendation algorithms in a certain way, improves the accuracy and diversity of recommendations, and then applies to more scenarios. There are seven commonly used mixing methods as follows[6]: Weighted: each recommendation algorithm is recommended and different weights are applied, and a single item is finally mixed and recommended to the user; Transformational: according to the standards designed by the designer, when the system encounters different standards, it selects one of the recommendation algorithms for recommendation; Hybrid: all the recommendation results of each recommendation algorithm are presented to users; Feature combination: the data generated by different recommendation algorithms are packaged together as different features, and then these features are used as the input of another method for recommendation; Serial: the first recommendation algorithm generates the recommendation result, and the second recommendation algorithm refines it successively, continuously screening and filtering, and finally obtains a relatively accurate recommendation result; Feature enhancement: the feature output of one recommendation algorithm is fed back to another recommendation algorithm as its input; Meta-level mixing: the model generated by the first recommendation algorithm serves as the input of the next recommendation algorithm.

LDA topic model is also called Latent Dirichlet Allocation (LDA), which was proposed by Blei et al in 2003. It is a probabilistic model of text set generation. It holds that there are multiple topics in a text, and each topic contains multiple words[7]. The LDA topic model is an unsupervised three-layer Bayesian structure, including documents, topics, and words. The model trains the topic model through some documents, then analyzes the topic of the target document, and finally outputs the topic of the target document in the form of probability. Given document set M and topic number K, the specific generation process of LDA can be divided into the following three steps[8]: (1) Dirichlet(α) randomly generates the corresponding topic distribution θ_m and the topic number $z_{m,n}$ from the polynomial distribution Mult(θ_m); (2) Dirichlet(β) randomly generates the word distribution φ corresponding to K topics, generates a word $w_{m,n}$ randomly from the polynomial distribution Mult($\varphi_{z_{m,n}}$) whose topic number is $z_{m,n}$, and generates a document by iterating N_m times; (3) Repeat this process until you have generated M documents on K topics.

III. Introduction to Algorithms

User text data in this article includes user dynamic content and chat history. The dynamic content posted by users and the chat records with other users can reflect the potential interests of users, so how to obtain the key information we need from these massive texts is the main task of this section.

When the topic or key words of an article are given, we can roughly understand its main content. Conversely, when we get an article, how to find out the representative information of the article and classify it? Naive Bayes classification is a good solution for many text classification problems, but it is powerless in the face of polysemy and polysemy problems. For example, "ink" refers to a liquid with pigment or dye, but can also be used to represent one's cultural knowledge. To solve this problem, the LDA topic model introduces the concept of topics between documents and words[9]. The LDA document generation process can be represented by an LDA graph model, as shown in Figure 3.

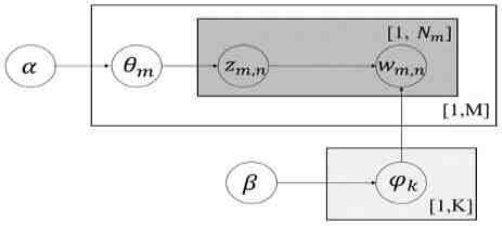

Figure 3 LDA diagram model

According to the generation process of LDA topic model, the joint distribution probability of all variables of LDA model can be expressed as:

$$p(\vec{w},\vec{z},\theta,\varphi|\alpha,\beta)=\prod_{k=1}^{K}p(\varphi_k|\beta)\prod_{m=1}^{M}p(\theta_m|\alpha)\prod_{n=1}^{N_m}p(z_{m,n}|\theta_m)p(w_{m,n}|z_{m,n},\varphi)$$

Joint distribution $p(\vec{w},\vec{z},\theta,\varphi|\alpha,\beta)$ is included in the observation variable to $\vec{w}$, hidden variables for $\vec{z}$, θ, φ estimates of these parameters, Gibbs sampling will be used in this paper. Gibbs sampling is a commonly used Markov chain Monte Carlo method, which is used to approximate sample sequence from a multivariable probability distribution when direct sampling is difficult. After several iterations of gibbs sampling, the probability distributions of θ and φ can be estimated as follows[10]:

$$\theta_{m,k}=\frac{n_m^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_m^{(k)}+\alpha_k)}$$

$$\varphi_{k,t}=\frac{n_k^{(t)}+\beta_t}{\sum_{t=1}^{V}(n_k^{(t)}+\beta_t)}$$

Specifically, the specific steps of the recommendation technology are as follows, and the algorithm flow chart is shown in Figure 4.

(1) Input user document D: merge the dynamic content of each user and all chat records of other users to form user document set D.

(2) Randomly select user U: randomly select a user U to

perform topic clustering for its user document du.

(3) Generate user topic distribution θ_{du}: Generate user U topic distribution using LDA topic model θ_{du}.

(4) Generate user's behavior feature p_u : The user's behavior feature is represented by the user's topic distribution. We believe that the larger the value of p_{ui}, the more obvious the user's behavior feature is.

(5) Calculate similarity: Calculate the similarity of behavioral features between target user U and all other users.

$$sim(u,v) = \frac{\sum_{i=1}^{K} p_{ui} * p_{vi}}{\sqrt{\sum_{i=1}^{K} p_{ui}{}^2} \cdot \sqrt{\sum_{i=1}^{K} p_{vi}{}^2}}$$

(6) Output top-n recommendation: The top N users with the largest similarity value form a top-n recommendation set and recommend them to the target user U.

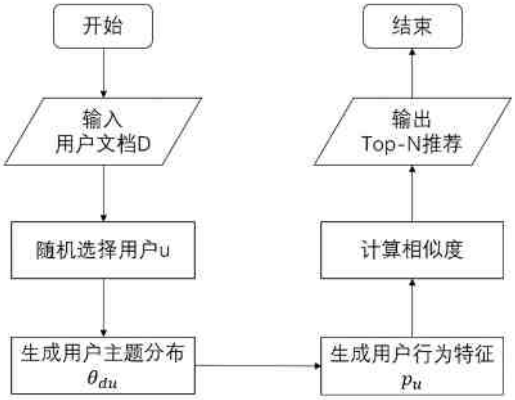

Figure 4 Flow chart of friend recommendation algorithm

IV. Conclusion

The algorithm aims to accurately recommend interested users to target users according to their chat files and dynamic content. In this paper, the development background of social network is introduced firstly. Then it introduces the related technologies of this paper, including recommendation algorithm and LDA topic model, an important model of friend recommendation algorithm mentioned in this paper. Recommendation algorithm introduces the main content-based recommendation algorithm, collaborative filtering recommendation algorithm and seven hybrid recommendation algorithms with different modes. Finally, a friend recommendation algorithm based on user text data is introduced, and each step is introduced one by one through the flow chart of the algorithm to achieve the target user's friend recommendation.

References

[1] Jing G , Gao M L , Xu B , et al. A hybrid recommendation algorithm based on social networks[C]// International Conference on Heterogeneous Networking for Quality. IEEE, 2015.

[2] Chen R , Hua Q , Chang Y S , et al. A Survey of Collaborative Filtering-Based Recommender Systems: from Traditional Methods to Hybrid Methods Based on Social Networks[J]. IEEE Access, 2018, 6:64301-64320.

[3] Bobadilla J , Ortega F , Hernando A , et al. Recommender systems survey[J]. Knowledge-Based Systems, 2013, 46:109-132.

[4] Wei S , Zheng X , Chen D , et al. A hybrid approach for movie recommendation via tags and ratings[J]. Electronic Commerce Research & Applications, 2016:83-94.

[5] Herbrich R , Graepel T , Stern D . Recommender System[J]. US, 2010.

[6] Schoeffmann K , Merialdo B , Hauptmann A G , et al. [Lecture Notes in Computer Science] Advances in Multimedia Modeling Volume 7131 ‖ Improving Item Recommendation Based on Social Tag Ranking[J]. 2012, 10.1007/978-3-642-27355-1(Chapter 17):161-172.

[7] YU, Hua; YANG, Jie. A direct LDA algorithm for high-dimensional data—with application to face recognition. Pattern recognition, 2001, 34.10: 2067-2070.

[8] WEI, Xing; CROFT, W. Bruce. LDA-based document models for ad-hoc retrieval. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006. p. 178-185.

[9] JELODAR, Hamed, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 2019, 78.11: 15169-15211.

[10] LU, Juwei; PLATANIOTIS, Konstantinos N.; VENETSANOPOULOS, Anastasios N. Face recognition using LDA-based algorithms. IEEE Transactions on Neural networks, 2003, 14.1: 195-200.