Review on State-of-the-art Technologies and Algorithms on Recommendation System

Haoyang Li[1], Yuanxu Wu[2] and Wei Xia[3]

[1]School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[2]School of Software Technology, Dalian University of Technology, Dalian, China

[3]Information Engineering College, Guangdong University of Technology, Guangzhou, China

Keywords: Hybrid filtering; Collaborative filtering; Sentiment analysis; Recommendation system; Evaluation.

Abstract. With the rapid development of Internet technology, we have entered into an era of information explosion. An obvious feature of this era is that it has a huge amount of data information. Facing with the huge amount of information, we need a system to effectively screen and filter the large-scale data. If a system can display the helpful information as much as possible, then users can save time to filter information. Therefore, how to design a system that can effectively screen important information and filter secondary information has become an important research topic in the era of big data. However, with the rapid development of the network and a large number of online information, there is a problem of information overload. A new way to solve the problem of information overload is to design and implement a recommendation system. The recommendation system can recommend information and products that interest users, according to user's information needs, interests, et al. In this paper, we introduce several recommendation systems based on different methods and algorithms, and we compare the effectiveness of assessments. The quantitative results of the user's emotional factors are applied into scoring matrix and similarity calculation in order to improve the effectiveness and robustness of recommendation system. When collaborative filtering based recommendation system combined with sentiment analysis, we can obtain an effective recommendation system.

1. Introduction

The rapid development of the network has a profound impact on enterprise development and personal life [1]. However, with the popularity of the Internet, the information is growing exponentially, which is far beyond what people are able to accept. Thus people may feel helpless when facing a large amount of information, this phenomenon is called information overload. When facing excess information, the boundaries between important and unimportant information are easily lost. How to quickly and effectively help users to obtain the important information has become a challenging work, and this topic is also the current academic hot issue.

In this context, recommendation systems are put forward. At present, the definition of the recommendation system is relatively vague, and there is a lack of authoritative definition in the academic circles. The widely quoted definition of recommendation system is proposed by Resnick et al. in 1997 [2].

Recommendation system has three components: recommend candidate, user, and recommended method. Users can provide personal preference information and recommendation requests proactively, otherwise, the recommendation systems can collect information proactively. Recommendation system can use different recommendation strategies to implement the process of recommendation. For instance, recommendation system can calculate the recommended results according to collection of personalized information and object data, and recommendation system feedbacks the results to the users [3].

At present, the e-commerce recommendation system has a good development and application prospect, and the research of the recommendation system is paid more and more attention, especially the personalized recommendation system. Almost all large-scale electronic commerce systems use recommendation systems with various forms, such as Amazon, CDNow, eBay, dangdang, et al. A successful e-commerce recommendation system can effectively retain users, improve sales and produce enormous economic benefits. At the same time, personalized recommendation system has become the hotspot issue in the field of electronic commerce.

2. Introduction Of Recommendation System And Algorithms

The characteristic of information retrieval is that the user's needs are different, whereas the retrieved information resources are relatively constant. The characteristic of information filtering is that the user needs are relatively fixed, whereas the information resources are constantly changing. Different with information retrieval and filtering, the characteristics of the information recommendation is the user's demand is not exact, thus we can only implement data mining process according to historical data and relate data. In recommendation system, the information resources are constantly changing, so the system needs to recommend information positively according to users' demand. For instance, in the e-commerce system, the new products are constantly provided, and the system should recommend which products is a typical problem of information recommendation. Recommendation algorithm is the core and key part of a whole recommendation system, to a large extent, the algorithm determines the type and performance of the recommended system. At present, there is no uniform standard for the classification of recommender systems, and many scholars have different classification of the recommended methods from different angles [4,5]. Mainstream recommendation methods basically include: content-based recommendation, collaborative filtering-based recommendation, knowledge-based recommendation.

2.1 Simple Description of Collaborative Filtering Algorithm.

Technology of collaborative filtering recommendation is one of the most successful techniques in the recommender system.

In typical collaborative filtering recommendation system, input data can usually be expressed as a user-item access matrix *R*, and it is denoted as $^{m\times n}$, where *m* is the number of users, *n* is the number of resources. Collaborative filtering based recommendation system can recommend information from the users' point of views, and this process is automatic. Namely, the system can recommend the information that interests users according to users' browsing history, thus users do not have to search recommended information for their own interest. Another advantage is that there is no special requirement for the recommended object, collaborative filtering-based recommendation system can handle unstructured complex objects, such as music, movies, et al.

Although, collaborative filtering recommendation technology has achieved great success in both research fields and applications, there are still some huge challenges in the study of the related research [6]. The first challenge is the recommended accuracy. For collaborative filtering algorithm, the sparsity of data is a great challenge. Because the user does not buy all the goods, so only a small part of products can be purchased. Therefore, the data matrix is very sparse, and data sparsity has also become the main reason that can influence collaborative filtering technology. The second challenge is scalability of collaborative filtering algorithm. Due to the sharp increase in the number of users and projects, collaborative filtering algorithm should be used in large-scale data in order to adapt to the changing data. Thus, the scalability of the algorithm is the main research problem in collaborative filtering algorithm. The third challenge is evaluation and measurement problems. Due to a large amount of evaluation and measurement methods, there is no definite criterion, so we have to continuously find out the effective way to better evaluate the performance of filtering algorithm.

2.2 Problem description.

Collaborative filtering algorithm can be described as follows. Assume we have a user set, denoted as $U = \{u_1, u_2, ..., u_m\}$; an item set, denoted as $I = \{i_1, i_2, ..., i_m\}$; and a set of preference degree, denoted as $R(t) = \{r_1(t), r_2(t), ..., r_m(t)\}$. Generally speaking, a user only buys a small part of the products, thus,

user-item matrix is very sparse. Collaborative filtering algorithm is designed to predict the user's interest in the products that have not been purchased yet. The general steps of collaborative filtering recommendation are described in figure 1.

Fig. 1 Collaborative Filtering Recommendation Flowchart

2.3 User-based Collaborative Filtering.

User-based collaborative filtering is an early collaborative filtering algorithm. Generally speaking, the early collaborative filtering algorithm is all based on users. Based on this assumption: if the scores of different users are similar, then other item scores are similar too. The basic idea of collaborative filtering is to look for other users u' that are similar to the current user *u*, and calculated the utility value $E(s,u)$ of object *s*. Finally the algorithm uses the utility value to implement the process of sort and weighting for all *s*. Collaborative filtering recommendation can recommend information to users according to the score which is similar to the nearest neighbor rating data [7,8]. The core part of the algorithm is to find the most similar nearest neighbor set for a target user who needs the recommendation service [9-12]. User-based collaborative filtering algorithms mainly have memory-based CF and model-based CF.

2.4 Memory-based Collaborative Filtering.

Memory-based collaborative filtering algorithm is similar to the lazy learning algorithm in machine learning. This method uses the whole user-item score data set to implement recommended process. The system uses statistical techniques to search for a group of users called neighbors who have the same historical preference with the target user. Generally speaking, the collaborative filtering algorithm based on users can be divided into two steps: neighbor selection and user preference prediction. In the process of neighbor selection, the system chooses the nearest neighbor according to the similarity degree of user's historical preferences. The scores of items *i* and *j* are respectively denoted as $\vec{i}$ and $\vec{j}$ in *m*-dimensional space:

$$sim(i,j) = \cos(\vec{i},\vec{j}) = \frac{\vec{i},\vec{j}}{\|\vec{i}\| \bullet \|\vec{j}\|} \quad (1)$$

Pearson correlation measure formula as follows:

$$sim(u,u') = \frac{\sum_{i \in I_{u,u'}} (R_{u,i} - \overline{R_u})(R_{u',i} - \overline{R_{u'}})}{\sqrt{\sum_{i \in I_{u,u'}} (R_{u,i} - \overline{R_u})^2} \bullet \sqrt{\sum_{i \in I_{u,u'}} (R_{u',i} - \overline{R_{u'}})^2}} \quad (2)$$

Where $I_{u,u'} = u \cap u'$, $R_{u,i}$ represents average preferences, $sim(u,u')$ represents the similarity degree of u and u'. Herlocker et al. put forward Spearman correlation function to select nearest neighbor. The calculation results of Spearman correlation function is similar to the results generated from Pearson function. For small-scale data set, Spearman correlation function will obtain better results[14].

Once the user's nearest neighbor is selected, the next step is to predict the preference for the specified item. In general, the predicted results can be calculated as follows:

$$P_{a,i} = \overline{u_a} + \frac{\sum_{b=1}^{N} (u_b - \overline{u_b}) \times sim_{a,b}}{\sum_{b=1}^{N} sim_{a,b}} \quad (3)$$

Similarly, once the nearest neighbor of the item is obtained, the preference of a user can be calculated by the user's nearest neighbor preference. There are many methods to calculate the prediction results, which can be divided into weighted category and regression category.

2.5 Model-based Collaborative Filtering.

Model-based collaborative filtering is to establish user's information as a model to predict user's preference. The operation speed and scalability of the algorithm can be improved in terms of the establishment of the model. For model-based collaborative filtering algorithm, the number of items and users has an important influence on the computational complexity and space requirements, because many users' preferences of electronic commerce are calculated online, so that a lot of algorithms cannot be applied into the actual environment. However, model-based method can solve this problem. A model can be built up offline, but it may take a few hours, a few days, even a few months. Once the model is established, it will be quick to predict the user's preference.

Bayesian network technology using the training set to create the corresponding model, and the obtained training model is very small, thus the application of the model is very fast. This method is suitable for the user whose interest is barely changed. There are a lot of model-based collaborative filtering algorithm that can be used to solve the problem of data sparsity. Ungar put forward a model which can implement the clustering process of users and items at the same time [15].

3. Recommendation Based On Text Sentiment Analysis

Text sentiment analysis is a work to detect, analyze and mine the preferences, opinions and emotions of users [16]. In the analysis of emotional tendency, we regard the neutral emotion as the reference point to analyze affective deviation [17], and we also analyze the deviation intensity in order to implement the quantification of results. Finally, according to the calculated values of emotion, the score matrix is constructed [18]. Subjective text with a subjective emotional color describes the author's ideas and views, whereas, objective text is a kind of objective cognition which is not emotional, and it has the characteristics of certainty and objectivity. Text classification concludes two parts: preprocessing and feature extraction. The text preprocessing stage contains the work of word segmentation and the removal of stop words, and this stage can be completed by using a word segmentation device, generally have relatively good results. Feature extraction is to obtain the characteristics of the text in the grammatical structure and semantic level, and uses these features to implement classification process. There are several common classification algorithms, K-means, artificial neural network, Bayes, decision tree, SVM, et al[19]. Many experimental results show that Bayes algorithm has a relatively good practical effect, and this method is one of the most successful classification algorithms.

The Bayesian method not only simplifies the model by assuming strong independence among the attributes, but also reduces the learning complexity [20]. For any class label *y* and attribute set *X*, we have:

$$P(X \mid Y = y) = \prod_{i=1}^{d} P(X_i \mid Y = y) \tag{4}$$

In order to avoid the situation that the posterior probability value is zero, we need to implement smoothing process, thus the improved posterior probability formula is:

$$P(t_i, c) = \frac{num(t_i, c) + 1}{\sum_{i=1}^{n} num(t_i, c) + V} \tag{5}$$

Where $P(t_i, c)$ is posterior probability, $num(t_i, c)$ is the number of occurrences of t_i.

After the process of preprocessing, we need to implement the process of sentiment analysis, and then complete emotional modeling work. Firstly, on the basis of the general sentiment dictionary, we construct domain emotional dictionary which is combined with corpus. Next, we choose the emotion words according to the emotion dictionary and mark the polarity of emotion words. Finally we can obtain emotional value of each review section and the entire review, and input this quantitative result to an evaluation matrix.

Extracting the emotion words and negative words in each comment segment, in this way we can calculate the emotion value according to polarity value of the given word in the emotion dictionary:

$$Polarity = \frac{\sum_{i=1}^{N} (-1)^k wp_i}{N} \tag{6}$$

Where *N* is the total number of emotional values, wp_i is the *i*-th polarity value. At the final stage, we construct the score matrix according to the emotion values, and we can use these emotion values to construct emotion vector in order to complete the modeling of users' emotion.

4. Evaluation of Several Recommendation Systems

In this section, we compare some recommendation systems proposed by several researchers in order to better evaluate the effectiveness and robustness of different algorithms. First we split the data set into train set and test set [21] with ratio of 4:1, and on this basis, we calculate the accuracy and recall rate. When cluster center *K* has different values, the sentiment analysis-based recommendation system will have different effects. The results of three recommendation algorithms are shown in Tab.1.

Table 1 Experiment result of three recommendation algorithms

Algorithm	**Value of *K***	**Correct recommended quantity**	**Total number**	**Accuracy rate%**	**Recall rate%**	**F1%**
Item-based	/	1076	1796	59.90	11.96	20.27
User-based	/	1103	1885	58.52	12.26	19.93
Sentiment analysis-based	10	1009	1964	51.34	11.21	18.38
	20	1082	1913	56.64	12.03	19.85
	30	1122	1850	60.61	12.47	20.67
	50	1110	1902	58.28	12.31	20.33
	100	1032	1922	53.55	11.45	18.87
	200	983	1838	53.45	10.93	18.11

Where the computational formula of are as follows:

$$\mathrm{Re}call = \frac{\sum_u |R(u) \bigcap T(u)|}{\sum_u |T(u)|} \tag{7}$$

$$\mathrm{Pr}ecision = \frac{\sum_u |R(u) \bigcap T(u)|}{\sum_u |R(u)|} \tag{8}$$

$$F1 = \frac{2 \bullet \mathrm{Re}call \bullet \mathrm{Pr}ecision}{\mathrm{Re}call + \mathrm{Pr}ecision} \tag{9}$$

The results shown in table 1, we can figure out that sentiment analysis-based recommendation algorithm has higher accuracy rate, recall rate and F1 with $K = 30$ than other algorithms.

From figure 2, we can figure out that, with the increasing of value *K*, the performance of sentiment analysis-based algorithm is first improved and then decreased. When $K = 30$, this algorithm can obtain the best performance.

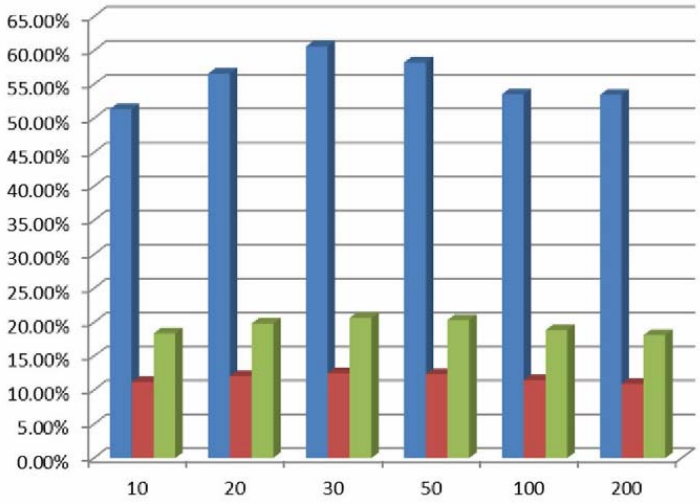

Fig. 2. Results comparison of performance indicators of the recommendation algorithm based on emotional tendency analysis with different values of K: blue: accuracy rate; red: recall rate; green: value of F1

References

[1] J. L. Herlocker, J. A. Konstan, L. G. Terveen, J. T. Riedl," Evaluation collaborative filtering recommendation systems," ACM Trans. Inf. Syst., 2004:5-53

[2] Resnick, Varian, Recommender systems[J]. Communications of the ACM, 1997, 40(3):56-58

[3] Breese J., Heckerman D. and Kadie C.. Empirical Analysis of Predictive Algorithms for Collaborative Filtering[A]. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, 1998:43-52

[4] Schafer JB,Konstan J, Riedl J. Recommender systems in e-commerce, In: Proc. of the 1st ACM Conf. on Electronic Commerce, Nwe York: ACM Press, 1999: 158-166

[5] Balabanovic M, Shoham Y. Fab: Content-based collaborative recommendation. Communications of the ACM, 1997, 40(3):66-72

[6] B.M. Sarwar, G. Karypis, J. A. Konstan, J. Riedl. Analysis of Recommender Algorithms for E-Commerce. In Proceedings of ACM E-Commerce 2000 Conf.

[7] Shardanand, U. Maes, P. Social information filtering: Algorithm for automating word for mouth. In Proceedings of the ACM CHI Conference on Human Factors in Computing Systems

[8] Konstan, J. Miller, B. Maltz, D. Herlocker, J. Gordon, L. Riedl, J. Gruoplens: Applying collaborative filtering to usenet news. Communications of the ACM, 1999

[9] Breese J. Hecherman D. Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[A]. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, 1998

[10] P Resnick, N Iacovou, M Suchak, P Bergstrom, J Riedl. An open architecture for collaborative filtering of netwews. In Proceedings of the ACM Conference on Computer Supported Cooperative Work, 1994

[11] Billsus D, Pazzani M. User Modeling for Adaptive News Access, UMUAL. Special Issue on Deployed User Modeling Systems.

[12] P Symeonidis, A Nanopoulos, A N Papadopoulos, Y Manolopoulos, Proceedings of the WebKDD Workshop held in conjuction with KDD, 2006, Aug.20-24, US

[13] Sarwar B M, Konstan J A, Borchers, A Herlocker,, J Miller. Using filtering agents to improve prediction quality in the GroupLens research collaborative filtering system. Paper presented at the ACM Conf. on CSCW

[14] Herlocker J. Konstan J. An algorithmic framework for performing collaborative filtering. Paper presented at the SIGIR.

[15] Ungar L H, Foster D P. Clustering methods for collaborative filtering. Paper presented at the Workshop on Recommendation Systems, 1998

[16] Xiang Yang, Chen Qian. A summary of text sentiment analysis[J]. Chengdu: Computer Applications, 2011, 32(12): 3321-3323

[17] Li Rongjun. Analysis of Chinese commodity review[D]. Beijing, 2011: 1-115

[18] Han Jiawei and Micheline K. Data Mining: Concepts and Techniques(2nd edition)[M]. San Francisco, California, USA: Morgan Kaufmann, 2006: 89-200

[19] Yang Y, Liu X. A Re-examination of Text Categorization Methods[C]. Proceedings 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIRp99), 1999, 42-49

[20] Luo Haifei, Wu Gang, Yang Jinsheng. Text classification method based on Bayesian[J]. Beijing, 2006, 27(24): 4746-4748

[21] Adomavicius G, Tuzhilin A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans on Knowledge and Data Engineering, 2005, 17(6): 734-749