

The Development of E-commerce Recommendation System Based on Collaborative Filtering

GuoFang Kuang^{1, a}

¹College of Information Technology , Luoyang Normal University, Luoyang,471022, China

^akuangguofang2012@163.com

Keywords: E-commerce; collaborative filtering; recommendation system

Abstract. The recommendation system in the e-commerce is to provide customers with product information and recommendations to help customers decide what to buy goods and analog sales staff to recommend merchandise to complete the purchase process. Collaborative filtering process is based on known user evaluation to predict the target user interest in the target, and then recommended to the target user. This paper proposes the development of E-commerce recommendation system based on Collaborative filtering. Experimental data sets prove that the proposed algorithm is effective and reasonable.

Introduction

In this context, the recommendation system (Recommender Systems) came into being, it is based on the characteristics of the user, such as hobbies, it is recommended to meet the user requirements of the object, also known as personalized recommendation system (the Personalized Recommender Systems). Recommended characteristics of the object, there are two types of recommendation systems, one based on the pages to Web data mining methods and techniques recommended object of the search system, for users to recommend web pages that meet their interests, such as Google; another online shopping (especially B2C type) environment, commodity recommended personalized recommendation system for referring users to meet the interests of the goods such as books, audio and video, said such a recommendation system for electronic Business personalized recommendation system, referred to as e-commerce recommendation system (recommender system in E-commerce).

Collaborative filtering recommendation system, as well as e-commerce website has been widely used, it is by far the most successful information filtering technology [1]. Collaborative filtering, also known as social filtering (Social Filtering), the basic idea of it is the degree of similarity by comparing user interests and behavior, to identify and target users with the same or similar interests and user groups, based on their resources evaluation to predict the target user's interest, to achieve the purpose of the recommended resources to the target user. This basic idea and is quite popular now word of mouth (word-of-mouth) is a bit similar. I believe we all understand, in real life, the most effective information is often recommended from friends.

Can easily provide a good recommendation for thousands of users, but for e-commerce sites often need to give hundreds of millions of users recommended, which on the one hand the need to improve the response time requirements, to provide users real-time collaborative filtering algorithm to be recommended; the other hand, should also take into account the storage space requirements to minimize the burden of the recommendation system running. This paper proposes the development of E-commerce recommendation system based on Collaborative filtering.

The research of Collaborative filtering

The starting point for collaborative filtering: with similar interests users might be interested in similar things. So, as long as the maintenance of data on user preferences, to analyze the draw users with similar tastes, then you can be similar to the views of customers to their recommended. Another possible starting point is: the users may be more similar to the preference to its purchase of goods. To

determine the degree of similarity between the goods according to the user evaluation of a variety of things, and then recommend those items closest to the user interest. The first idea to the relationship between the client and the client as the center, it is in order to a train of thought while items with the relationship between the focus of attention.

Collaborative filtering process is based on known user evaluation to predict the target user interest in the target (ie, evaluation values), and then recommended to the target user [2]. Collaborative filtering system applied algorithm is in progress to user interest prediction rules to follow in this process, more consistent rules and the actual law, the more accurate prediction of user interest, filter information, the better is shown by equation 1.

$$I_{k-1}(X) = I_k(X') = I_k(W(X, P)) \quad (1)$$

Sparseness (sparsity) is one of the recommended techniques. Implementation of collaborative filtering technology first need to use the user - of the evaluation matrix of user information that, although this is very simple in theory, but in fact, many e-commerce recommendation system to a large number of data processing in these systems general users to buy the total amount of goods accounted for about one percent of the site's total amount of commodities, and therefore the evaluation matrix (user - matrix) is very sparse. In such a large amount of data and sparse, on the one hand, difficult to find the nearest neighbor sets of users, on the other hand the cost of similarity calculations will be great.

By the nature of the cluster shows the nearest neighbors of the target users are distributed in the clustering of the highest similarity with the target user, so there is no need to check the nearest neighbors of the target user in the user data space, only with the target user can query to the highest similarity with the clustering of most of the neighbors of the target user, the highest clustering similar to the target user for the entire data space, the search space is much smaller, so the proposed method can improve the online nearest neighbor search speed, and effective real-time requirements to meet the recommended system, as is shown by equation2.

$$\frac{\Delta x_i(k)}{\sigma_i} = [x_i(k+1) - x_i(k)] \sigma_i^{-1} \quad (2)$$

Through the experimental data sets in Each Movie on the above three methods are compared with the above three rights were re-used in the calculation of the Pearson correlation coefficient to calculate the correlation coefficient, and then substituted into the prediction formula to calculate the results show that the recommended some improvement in the quality of the results [3]. Each Movie data sets collected between 1996 ~ 1997 72916 users 2456676 1682 film and evaluating values of the matrix is very sparse. For convenience, the experiments of any take 10000 records and users of these records at least 20 films have been evaluated, the data set is divided into a training set (8000 users) and test set (2000 users), the evaluation criteria used in the experiment the mean absolute error MAE to evaluate the accuracy of the forecast, the smaller the value, show that the prediction accuracy of the higher. The results are shown in Figure 1.

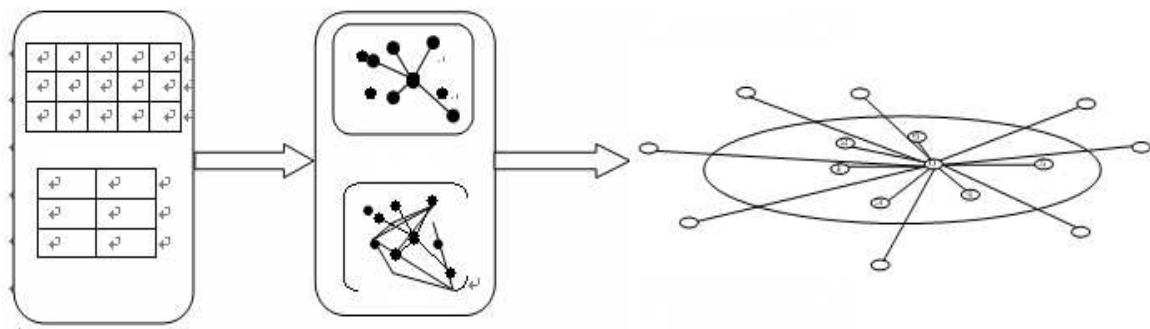


Figure. 1 The result of several collaborative filtering algorithm performances

Nearest neighbor queries based on user clustering only the nearest neighbor of the target user's query closest to the target user clustering, it can only guarantee that the search for most of the neighbors of the target user, and can not guarantee that all the nearest neighbor search to the target user's make recommendation system accuracy of a certain reduction.

Such algorithms, the most widely used is based on the neighbor method (Neighborhood-based methods) or called correlation-based method (correlation-based), by selecting the target user with a similar user sets this similar user focus on the user a target of evaluation, the formation of the user on the target items of interest measure and predict. The implementation of this algorithm can generally be divided into three steps:

- (1) Calculate the similarity between each user;
- (2) According to the degree of similarity between the users to select a similar subset of users;
- (3) By a particular method, similar to the user to focus on user evaluation of the target to form the prediction of the target users interested.

The advantage of this algorithm is the following: 1) fully take into account that part of the user of inconsistencies between the evaluation value of the neighbor user, the user evaluation value changes; 2) to avoid the error caused by the evaluation value of most users are too concentrated nearest neighbor, because the number of these values often than the other nearest neighbor users especially critical nearest neighbor users have a greater impact, resulting in the deviation; 3) for the new user preferences to time according to the judge added to the user focused, as is shown by equation 3 [4].

$$\overline{I_k} = \frac{\sum_{x \in \Omega_i} [I_k(X)]}{L \cdot L}, \overline{I_{k-1}} = \frac{\sum_{x \in \Omega_i} [I_{k-1}(X)]}{L \cdot L} \quad (3)$$

More collaborative filtering in the analysis of the problems facing the implementation of the recommendation system, they have in common is to take into account the formation of the nearest neighbors (including user access to information sufficient to calculate consuming, etc.). But you should see the collaborative filtering recommendation system implementation, to obtain the nearest neighbor users, must be calculated by a certain similarity between users, and then determine the optimal number of neighbors, the formation of the neighbor sets of users. The similarity is calculated as follows equation 4.

$$k_E(x) = \begin{cases} \frac{1}{2} c_d^{-1} (d+2)(1-x^2), & \text{if } x < 1 \\ 0, & \text{if } x \geq 1 \end{cases} \quad (4)$$

To better solve the data sparseness of collaborative filtering recommender system implementations, synonyms (similar products with a different name to describe and can not find this correlation), and the present use is widely used in information retrieval dimensionality reduction - latent semantic indexing (Latent the Semantic the Indexing, LSI), used to solve the problem of synonyms and policemen.

In order to improve the efficiency of the recommended online user's offline processing section complete the class within the similarity calculation, the calculation of similarity coefficients using the cosine similarity, that is calculated using the formula (4). Calculate the similarity coefficients between users within the class; it is the calculated results are written to the table SimiCoefficient. Page recommendation, it is first to obtain the target user IP, and then determine that the user is the first visit or have visited the site. If the user is the first visit, choose to access the high frequency of the first N terms as the recommended content.

The development of E-commerce recommendation system based on Collaborative filtering

E-commerce website in change for the purchase: the visitors of e-commerce system in the process often does not desire to buy, the personalized recommendation system to recommend to the user they are interested in goods, which led to the purchase process. Cross-selling capabilities to improve

e-commerce site: personalized recommendation system to provide users with the users to buy other valuable commodity recommendation, they really need in the purchase process did not expect the goods purchased in the list of users from the system recommended in order to effectively cross-selling e-commerce system.

Collaborative filtering algorithm, calculating the similarity between the user and recommended into a process, in order to improve the speed of the recommended paper calculating the similarity between users on the offline processing section to reduce the online recommended amount of computation [5]. Offline processing part of the data pre-processing and user clustering has been introduced in the previous section, the following collaborative filtering recommendation part, recommended part of the program, as is shown by equation5.

$$\Phi(m) := \left[\left(\prod_{i=0}^0 A(mM + i) \right)^T, \left(\prod_{i=1}^0 A(mM + i) \right)^T, \dots, \left(\prod_{i=M-1}^0 A(mM + i) \right)^T \right]^T \quad (5)$$

Due to a variety of recommended methods has advantages and disadvantages, so in practice, the combination of recommended (Hybrid Recommendation) is often used. Research and application is a combination of content recommendation and collaborative filtering recommendation [6]. The easiest way is to use content-based approach and collaborative filtering recommendation method to generate a recommended forecast results, and then use a method to combine the results. Although in theory there are many recommended a combination of methods, but not necessarily valid in a particular issue, one of the most important principles of the combination recommended by the combination to be able to avoid or compensate for the weakness of their respective recommended techniques, that is the nature of equation 6.

$$\|\overline{V(k,t)}\| = \frac{1}{n} \sum_{i=0}^{n-1} \|V_i(k,t)\| \quad (6)$$

Close neighbor of the user based collaborative filtering technology is relatively common, the core issue for the current user to find the k most similar neighbors to predict the user's interest. This method is difficult to solve problems encountered in practice, first, sparse, means the initial period of the system, system resources has not been enough evaluation, and it is difficult to use these evaluations to find similar users. The second is scalability, that is, with the increase of users and resources, the method performance is getting worse.

(1) Each t_i is clearly visible, where T can be represented as a collection of the following form: $t_i = \langle w1(t_i), w2(t_i), \dots, wn(t_i) \rangle$;

(2) First calculate the similarity between the current user and other users can be between users of one or a few related features;

(3) Randomly selected from T_i Initial_Size number of users to the formation of the initial user set T'_i ;

(4) We assume to predict the evaluation value of the item i , v_u, i said that user u on item i and evaluating values of T_i in item i is expressed as V_i , the user u , a collection of other items of the evaluation value is expressed as $F(u, i)$ a description of the user u set.;

(5) randomly retrieve a value from a discrete $[0, 1, \dots, (2^k - 1)]$ set of integers, denoted by r , the re-election after the time-piece.

(6) For (int $i=1, i \leq k, i++$)

(7) $U'_k = \theta(U_k), S = [], S' = [], S = S \cup U_k, S' = S' \cup U'_k, V = U_k$;

(8) If to be recommended only belong to one scene, derived from step (3) predictive value is asked for the results. With recommended items belong to more than one scene, a few scenes predictive value of the average demand for the results.

In addition to real-time analysis algorithm, this also added time the weight of this paper, collaborative filtering algorithm with the traditional collaborative filtering algorithm based on user recommendations accuracy. In this study, we hope that the final algorithm is able to accurately predict the evaluation project evaluation points for the user to make a more accurate recommendation algorithm accuracy (Precision) is to evaluate a major indicator of the recommendation algorithm. The practice of most of the literature, we forecast and actual values of the mean absolute error (MAE: Mean Absolute Error) as a standard measure of algorithm accuracy.

In order to verify the algorithm in this paper, experimental data sets by using a widely used and recognized, the proposed algorithm with conventional algorithms in the interest degree forecast accuracy to prove that the proposed algorithm is effective and reasonable, as is shown by figure2. Used in this study the machine configuration Intel Pentium 4 processor, 1G memory, 300G hard disk, the operating system to Windows XP, use the tools of MATLAB 7.

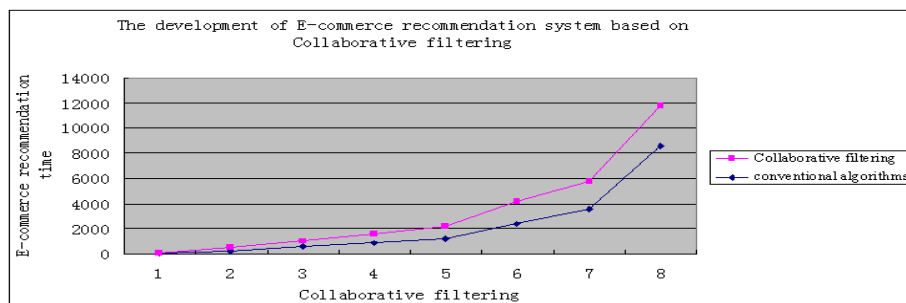


Figure. 2 The compare of E-commerce recommendation based on collaborative filtering with conventional algorithms

Summary

Collaborative filtering recommendation system, as well as e-commerce website has been widely used, it is by far the most successful information filtering technology. The recommendation system is the use of statistical and knowledge discovery techniques to solve the provision of goods recommended system with the goal of customer interactions. This paper proposes the development of E-commerce recommendation system based on Collaborative filtering. In order to verify the algorithm in this paper, experimental data sets by using a widely used and recognized, the proposed algorithm with conventional algorithms in the interest degree forecast accuracy to prove that the proposed algorithm is effective and reasonable.

References

- [1] Sarwar B, Karypis G, Konstan J, Riedl J. Analysis of recommendation algorithms for E-commerce. In: ACM Conference on Electronic Commerce. 2000. 158~167
- [2] Sarwar B, Karypis G, Konstan J, Riedl J. Item-Based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International World Wide Web Conference. 2001. 285~295
- [3] Lee, W.S. Online Clustering for Collaborative Filtering, National University of Singapore, School of Computing Technical Report TRA8/00, 2000
- [4] Schafer J B, Konstan J A and Riedl J. E-Commerce recommendation applications[J]. Data Mining and Knowledge Discovery, 2001, 5(1-2):115-153.
- [5] R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web data. PhD thesis, Dept. of Computer Science, University of Minnesota, May 2000
- [6] Mobasher B, Cooley R, Srivastava J. Automatic personalization based on web usage mining[J]. Communications of the ACM, 2000, 43(8):142-151