User clustering model based on collaborative filtering

Shaolin Wang [a], Laisheng Xiang [b] and Xiyu Liu [c]

School of Management Science and Engineering, Shandong Normal University, Jinan 250014, China;

[a]wangshaolin1116@163.com, [b]xls3366@163.com, [c]sdxyliu@163.com

Abstract

With the rapid development of Internet technology, the overall number of network information showing explosive growth, the growth of data in more comprehensive information for users at the same time, the shortest period of time to obtain the information users need has been plagued by users and website management, data obtain become main problems that puzzled the Internet users. In this environment, the application of the recommended system provides a good solution to these problems. Collaborative filtering recommendation method provides a very convenient service for users in the application of e-commerce. However, the rapid growth of network information, lead to collaborative filtering recommendation with a wide variety of defects, seriously restricting the recommended efficiency. User clustering is recommended to improve efficiency and user satisfaction new breakthrough, is the traditional recommendation algorithm based on user clustering of innovation of the algorithm in this paper, the algorithm uses the user browsing the web in the feedback of implicit data, calculating similarity of users, in order to achieve similar user clustering, based on Clustering Based on the target user according to a recent neighbor implicit preference was calculated, the algorithm not only reduced the looking for the most similar to the user's computing dimension, and recommended speed a breakthrough progress.

Keywords

Personalized recommendation, user clustering, collaborative filtering.

1. Introduction

The rapid development of the Internet makes it easier for users to create information and share information, leading to exponential growth in the amount of information on the web, and it is well known that in the face of so much information, the early portals of the web, through the way of categorizing catalogs, show different types of resources But this approach requires a user-level search. However, with the rapid development of web2.0 , the growth of network resources has been far beyond people's imagination, the way to find resources through the classification of navigation trees becomes extremely cumbersome and inefficient, in this trend, Baidu, Google and other companies seek search engine way to find the resources and information users want, but this is the search engine does not fully meet the needs of users, especially when the user does not have a clear search target and the content of the need is not very good language description, the search engine will lose the original advantages of efficiency, And the content of the search is easily affected by the "Matthew Effect", and the popular search terms and search results are presented in front of the search results to get more attention and clicks, But compared to the results of the unpopular will get less attention and clicks. For these reasons, the traditional search engine can not be from such a large amount of information to get users interested in the content, in such an environment, the recommendation system began to be more and more used in a variety of web sites, to provide users with services to help users get the necessary information from the complex .

As a core component of electronic commerce, the recommendation system has a good chance of development in various fields. For example, the information recommendation site can provide users with the customization of the content of the network, E-commerce sites can recommend their users

interested in goods, Customer center and counter services can provide users with personalized services. 1997 Year,Resnick and other people for the first time the definition of e-commerce recommendation system: "It is based on E-commerce Web site as a platform for consumers to provide goods information and advice to help them decide what products should be purchased, The simulation salesman assists the consumer to complete the purchase process and personalized recommendation systems need to be able to provide recommendations not only to customers, but also to attract new customers, retain old customers, and generate huge profits.

Recommendation algorithm as the essence of recommendation, its development has been more and more attention by the academic community, now generally simple to classify the recommendation algorithm as follows: Based on association rules recommendation algorithm, content-based recommendation algorithm and collaborative filtering based recommendation algorithm. As the core component of e-commerce, the recommendation system has been well developed in every field. Collaborative filtering recommendation system in various fields of large-scale use, to a certain extent, user-friendly. But the recommendation system is not always the same, but also with the user's increase and information growth dramatically change, accompanied by a variety of deficiencies. Among them, recommended system accuracy, recommended system scalability, and new users and new project entry will be difficult to get recommendations, that is, collaborative filtering cold start problem. These problems have seriously restricted the improvement of recommendation efficiency and user experience.

This research has made some improvement on the basis of collaborative filtering, which makes use of user's recessive preference and user clustering as the innovation of this research. Through the research of the user recessive data, it is not difficult to find that the recessive data can find the user's interest more accurately than the dominant data. And user clustering can be a good solution to the number of information and user growth caused by deficiencies. This paper is the research description of the algorithm of user clustering, and at the end of the paper, we make a summary of the clustering method.

2. Collaborative filtering

Collaborative filtering is done by collecting user behavior information when browsing the site, and then identifying users with the highest similarity to current user interests, and offering their preferred items to the target user for their choice, which is the TOPN recommendation. The recommendation technology based on collaborative filtering can be divided into three parts: Data mining processing, determining the closest neighbor, generating recommendation data. At present, most of the recommended algorithms used in e-commerce Web sites are mining users ' preferences through explicit feedback. For example: User browsing rating data, users in the site registered account interest preferences.

2.1 Similarity Algorithm

Computing the similarity of the user is a very important step in the system filtering algorithm, according to the degree of similarity among the Web surfers, to determine the interest set of the closest neighbor to the recommended target, and to give the information of interest to the target user. Commonly used three similarity calculation methods are: cosine similarity, correlation similarity and modified cosine similarity.

2.1.1 Cosine Similarity

To put it simply, cosine similarity uses the cosine of two vectors in vector space to measure the difference between two individuals, the larger the cosine, the smaller the angle between the two vectors, and the smaller the difference between the individuals, the higher the similarity of the two vectors. In the recommendation system, we use this algorithm to bring the two user's scoring data in, which computes the user similarity, assume that the user and the similarity of $\mathrm{Sim}(u_1, u_2)$.

$$\mathrm{Sim}(u_1,u_2)=\frac{\sum_{k=1}^{n} R_{u_1,i}\times R_{u_2,i}}{\sqrt{\sum_{k=1}^{n} R_{u_1,i}^{\ 2}\times\sum_{k=1}^{n} R_{u_2,i}^{\ 2}}} \tag{1}$$

Formula (1), $R_{u_1,i}$ on behalf of the user u_1 for the score of the object i, $R_{u_2,i}$ On behalf of the user u_2 for the score of the object i .

2.1.2 Correlation Similarity

Correlation similarity is through Pearson correlation coefficient (Pearson correlation coefficient) The calculates user affinity. First of all, we can analyze the score information of two website visitors on the net, through the collation analysis can get their common score of the same project set, calculate the two vectors of the peason correlation coefficient, and in order to calculate the similarity of two users use the formula below to calculate the similarity between users u_1 and u_2 .

$$\mathrm{Sim}(u_1,u_2)=\frac{\sum_{i\in I_{u_1,u_2}}(R_{u_1,i}-\overline{R_{u_1}})(R_{u_2,i}-\overline{R_{u_2}})}{\sqrt{\sum_{i\in I_{u_1,u_2}}(R_{u_1,i}-\overline{R_{u_1}})^2}\sqrt{\sum_{i\in I_{u_1,u_2}}(R_{u_2,i}-\overline{R_{u_2}})^2}} \tag{2}$$

2.1.3 Fixed cosine similarity

Because the cosine similarity algorithm does not take into account the different evaluation standards between users, resulting in the rating of the user interest is not uniform, so that the use of this algorithm to calculate the deviation of the user similarity will be very large. Fixed cosine similarity can solve this problem to some extent, introduce mean value, use different user's score minus mean value to calculate the user similarity.

$$\mathrm{Sim}(u_1,u_2)=\frac{\sum_{i\in I_{u_1,u_2}}(R_{u_1,i}-\overline{R_{u_1}})\left(R_{u_2,\mathrm{i}}-\overline{R_{u_2}}\right)}{\sqrt{\sum_{i\in I_{u_1}}(R_{u_1,i}-\overline{R_{u_1}})^2}\sqrt{\sum_{i\in I_{u_2}}(R_{u_2,i}-\overline{R_{u_2}})^2}} \tag{3}$$

$I_{u_1,\ u_2}$ represents users common scoring Project set, $R_{u_1,i}$ And $\mathrm{R}_{u_2,\ \mathrm{i}}$ represent the users scoring for the project i, $\overline{R_{u_1}}$ And $\overline{R_{u_2}}$ represent the average of the score for the item.

2.2 Collaborative filtering challenges

The popularization of the network and the rapid development of information technology, the constant expansion of the data scale will inevitably lead to the traditional user-dependent scoring information on the collaborative filtering of the recommended way gradually revealed a variety of defects, such as the more common cold start, data scarcity, the recommendation of the system scalability, special user needs. These common defects severely reduce the efficiency of the personalized recommendation system, which ultimately leads to users not being able to get the recommended content or recommending to the user the items they dislike.

2.2.1 Scarcity of data

Get the preferences of Web site visitors is not so easy to imagine, first of all, users need to enter the account when the initiative to fill in the personal interests of the relevant information, but also need to obtain their information on the evaluation information, belong to the passive access to information, the way there are various limitations. Real life, users of the initiative on the project scoring and fill in

interest hobby information is not high, users face a very large number of projects, it is not possible for each object can be evaluated, and evaluation of the project to effectively evaluate only a small part of all items, less than the entire product of the 1percent, This makes the user - object matrix data missing. The system cannot obtain sufficient data to demonstrate the accuracy of its calculations, so it is not possible to determine the reliability of close neighbor users.

2.2.2 Cold start problem

Cold start can be regarded as the extreme phenomenon of data scarcity in some sense, and the traditional collaborative filtering recommendation method is based on existing users. also has information about the target user's recommendation based on the interest of the already-owned viewer, but when the new content or project first appears on the E-commerce site, no visitor evaluates and scores it, This system will not be able to get the existing visitors to the new content or object evaluation information, naturally can not get the opportunity to recommend to target users. When a new visitor has just entered the system, it is similar to this, because the new user has not yet evaluated the content information, the system cannot obtain the user's preference information through the grading data. Therefore, the recommended content of the recommendation system is unreliable for new users.

2.2.3 Special User issues

Depending on the user rating may be the most users of the recommendation results, but in real life, there may be a small number of users with special preferences, the recommendation system does not have much effect. In the recommendation system of e-commerce, the need for similar users to recommend a prerequisite, that is, the majority of users of the interests of the performance of a more uniform, but this prerequisite applies only to most people, once the target user's needs and ideas are different, when they get close neighbor users, there will be deviations, The result of the recommendation is incorrect.

These problems have seriously affected the recommended accuracy rate and user experience with the increase of network information. Based on this research, the traditional recommendation algorithm is improved, and the original user clustering is improved.

3. User clustering based on implicit data

In general, a faster system is recommended, its accuracy may be unsatisfactory, how to improve the content of the recommendation of accurate consideration, fast and scalable good algorithm becomes our urgent need to solve the problem, the study has improved the user clustering algorithm, first of all, the most basic is to find the user clustering indicators. Most of the recommended algorithms are based on the user's dominant information collation analysis of data indicators, in the large number of data indicators are summarized and calculated to feedback the user's preferences. But there is an unavoidable problem, that is, the large number of dominant information in the data processing time consuming, and then there is the acquisition of these data requires the cooperation of users, is passively obtained. Once the user in the registration of less information, or did not evaluate any object, then the recommendation for the user is not to start, seriously affecting the quality of recommendations, resulting in the loss of a large number of users.

Recessive data refers to the hidden information of user's browsing record, the stay time, the purchase record, the collection record, the click Amount and so on on the website, can analyze the latent interest point of users, and the recessive data can reflect the user's preference better than the large user dominant data. Moreover the current data obtains needs the user to cooperate, belongs to the passive obtains, but the invisible data is all the generation, as long as the user browses the website to leave the corresponding information, this has brought the very great convenience for our research. Through these recessive data index, the user's similarity is obtained by substituting the formula, and the efficiency of the recommendation system is improved effectively.

As the core step of this research, user clustering occupies an irreplaceable position in the recommendation system. After obtaining the user's stealth index, the next step is to the user to cluster

operations, clustering is generally divided into the following three steps: Data mining, similarity calculation, clustering algorithm.

3.1 Data Mining

User behavior is obtained mainly through the front end of the page to insert code, you can collect user data information, and then get the user's behavior log. After getting the user's behavioral log, a lot of the log content is not what we need, the real need is only a small fraction of the content, which requires filtering and data processing.

Web Content Mining by obtaining a large number of Web logging, discovering the behavior patterns of users accessing Web pages, predicting user browsing behavior, the end result of theWeb log mining is often to get user preferences. By analyzing the user's log, we can get the user's stay on the site, purchase records, collection records, clicks, User browsing paths and other information, but this information these behavioral logs are difficult to measure the user's interest in an object, the value of this study is limited. On the other hand, these hidden data are not as easy to use as the user's score, and the resulting behavioral data need to be processed.

3.2 Data Processing

In this study, the user's operation information in a specific page is separated from the user's browsing path, which reflects the user's invisibility preference on the Page object by the user's behavior log in the current page. When the user's network browsing path is analyzed globally, the similarity of user preference information obtained by the current page feedback is similar to the user's path similarity, and the similarity of users is calculated according to their weights.

3.2.1 Data processing for specific page objects

According to the user's behavior log, delete the useless data, get the user behavior information between the user - object, in which the user's stay time on the website, whether the user collects the object, the user clicks and the user browsing path four kinds of information can be as the recessive data of this research. The three-dimensional coordinates of the user's stay in the Web page, the user's collection, number of clicks, or the frequency of the mouse movement three kinds of information are set to represent the user and project preferences.

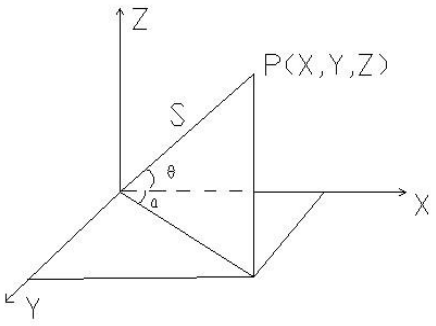

Figure 1

As the X axis in Figure 1 Indicates whether the user is a favorite, theY axis represents the number of clicks by the user or the frequency of the mouse movement, and the Z axis represents the user's stay time. The midpoint of the figure P (x,y,z) can be abstracted as a user, and S is the distance from the point p to the origin, which can be used to indicate the user's preference for the project.

In order to achieve the above prediction of user preferences, the need for the arrival time, the number of clicks, whether or not to collect information data processing, to obtain easy access to the use of the formula can be calculated:

Assuming we have the behavior information of the user U, we can analyze the parameter information for user U:

Whether the user collects parameters: IsCo=$\begin{cases}1\\2\end{cases}$ (where 1 indicates that the user did not collect the information, and 2 indicates that the user has collected the information).

User's Click quantity parameter: Clicks = $\begin{cases}3\times C_u\left(C_u > \overline{C}\right)\\2\times C_u(C_u = \overline{C})\\C_u(C_u < \overline{C})\end{cases}$, C_u represents the amount of clicks and mouse movements that a user U accesses when accessing the object (units: Times / minutes), $\overline{C}$ represents the average number of hits for all users. When the u hits more than the average, the user is more interested in the information.

The same can be based on the average stay time compared with the user to get the user stay time (unit: seconds) of the parameters:

Time=$\begin{cases}3\times T_u\left(T_u > \overline{T}\right)\\2\times T_u(T_u = \overline{T})\\T_u(0 < T_u < \overline{T})\end{cases}$

Because the user's behavior reflects the user's preference degree of effective form inconsistent, reflected to the reality is that when the user browses to a certain information, its stay time is longer, also added a collection, but because the information itself reason users do not need too many clicks or mouse movement, Therefore, in order to more accurately reflect the user in the establishment of the three-dimensional preference information, the three types of information can be standardized processing. The following formulas are introduced to calculate the position of the user in the three-dimensional coordinate system.

$P(X,Y,Z)=\left(IsCo\times a,\ Clicks\times b,\ Time\times c\right)$ (a+b+c=1)

Where a, b and c are used to adjust parameters to determine whether the collection, number of clicks, and the length of stay reflect the user's implicit preference for content in the current page.

3.3 Similarity Calculation

The second step in the user clustering is to compute the user similarity. There are many ways to compute user similarity, there are Euclidean distance similarity, Manhattan distance,spearman rank correlation coefficient algorithm , Tanimoto The factor (also known as Jaccard factor).

For the data that is available in the section 3.1, you can compute the user's similarity by using distances between two points in space:

$$sim_1(u_1, u_2)=\text{distance}(u_1,u_2)=\sqrt{\left(X_{u_1}-X_{u_2}\right)^2+\left(Y_{u_1}-Y_{u_2}\right)^2+\left(Z_{u_1}-Z_{u_2}\right)^2} \quad (4)$$

According to the above distance calculation can reflect the similarity between users to some extent, but this kind of calculation is based on the absolute value of each dimension feature, so we need to ensure that each dimension index is at the same scale level. Because of the different standards of user collection, stay time, number of clicks, or mouse movement of three kinds of data, there may be a significant deviation between the results and the actual results.

On this basis, we find that the application of a class of important information in the user's Web behavior Log can improve the result of computing user similarity, that is, the user's browsing path. The relationship between users can be indirectly reflected by the similarity of the

paths. For example, when the user's browse path order is not considered, the user can be calculated with the following formula

$$sim_2(s_{u_1}, s_{u_2}) = \frac{|s_{u_1} \cap s_{u_2}|}{|s_{u_1} \cup s_{u_2}|} \tag{5}$$

s_{u_1}, s_{u_2} represents path information of two users, $|s_{u_1} \cap s_{u_2}|$ represents the same number of paths for two users. $|s_{u_1} \cup s_{u_2}|$ represents the total of users. Although the above formula reflects the similarity of user browsing path to a certain extent, but the user browsing the page order will affect the accuracy of the algorithm, so considering the different page order, the following improvements are made:

$$sim_3(s_{u_1}, s_{u_2}) = \frac{|com(s_{u_1}, s_{u_2})|}{\max(s_{u_1}, s_{u_2})} \tag{6}$$

which $|com(s_{u_1}, s_{u_2})|$ represents the same path length for two users browsing the page order, $\max(s_{u_1}, s_{u_2})$ represents the maximum length of a two user browsing page path.

To synthesize the above three kinds of cases, compute user similarity can use the following algorithm:

$$sim(s_{u_1}, s_{u_2}) = \alpha \times sim_1 + \beta \times sim_2 + \lambda \times sim_3 \tag{7}$$

α、β、λ is adjust parameters, determine the size of the proportion of users in the calculation of the similarity.

3.4 User Clustering

The characteristics of the Internet determine that the network is always generating a large amount of information, the traditional collaborative filtering in the processing of large amounts of data gradually showed weakness, user clustering to make up for the lack of collaborative filtering. The recommendation algorithm based on user clustering is based on the visitor's Invisible Access index data, and the similarity of the visitors is clustered into a user set. Focus the same class on different people who have the most similar preferences, and when the target user visits the E-commerce site, first determine which cluster he belongs to, and make recommendations based on the preferences of the closest neighbor user in the same cluster. This reduces the computational complexity of looking for neighbor users and improves system efficiency.

According to different scenarios, requirements, in the application of clustering algorithms are similar, and at present, the most widely used clustering algorithm is also used K- means algorithm, K-means The algorithm is a typical clustering method based on partitioning, which is widely used in various fields. This clustering algorithm is simple, easy to control, fast, and can handle large data integration as its most important advantage. However , the K-means algorithm also has its unavoidable drawbacks: The number of cluster centers is difficult to determine, the clustering results are unstable, when large data volume is encountered, the time overhead is large issues such as are also a constraint to their use. To address these issues, the literature gives a k-means SCANalgorithm, which is an improved The K-means algorithm does not result in different clustering results because of the random determination of the initial cluster center. and solve the problem that the number of clustering is difficult to determine, in the literature also proposed by dynamically setting the cluster size, so that dynamic determination can be implemented K-means the number of clustering in the. But these improved algorithms also have difficulty in distinguishing whether the clustering is normal. In this paper , the algorithm is improved to adapt to the research environment for the defects of K-means.

3.4.1 K-means algorithm :

Input: User preference item set, parameters K ;

output: Calculated user Set K ;

Step 1, retrieve n items for user preferences, assuming that all projects make up a project set $I = \{i_1, i_2, i_3 \ \ldots \ i_n\}$.

Step 2, retrieve all participating m users, assuming that the user set $U=\{u_1, u_2, u_3 \ \ldots \ u_m\}$.

Step 3, select the K user from the m user as the cluster Center for the initial state, set to $W = \{w_1, w_2, w_3 \ \ldots \ w_k\}$. According to the nearest distance principle, user u_m $(u_m \in U)$ participates to the cluster center w_k $(w_k \in W)$.

Step 4, redefine the cluster center based on the average of the K user clustering

$$C=\{c_1, c_2, c_3 \ \ldots \ c_k\}, c_k (c_k \in C)=\frac{\sum_{i=1}^{n_k} u_i}{n_k}$$

which n_k represents the number of users in the k cluster.

Step 5, repeat steps 3, step 4until the cluster center no longer has a huge fluctuation or the number of users of the cluster is relatively stable.

3.4.2 Improved k-means algorithm :

Step 1, determine the number of clustering and cluster centers: to resolve the problem of K-means algorithm is difficult to determine the cluster center, redefine the distance by formula (5).

Step 2, user clustering: In step 1 , we have identified the number of clusters and cluster center set C, we need to take advantage of the formula (7) to compute the similarity between the target user and clustering center c_k . In the same cluster, the closest neighbor with the target users, content is recommended.

In this paper, the description of the user clustering algorithm improves the limitation of the system filtering algorithm to some extent, especially in solving the typical problem of collaborative filtering. Compared with the traditional K-means in the initial stage of user clustering, the clustering algorithm of this study can effectively reduce the problem that the final clustering result caused by the random setting of cluster centers is very different and the result is unstable.

4. Defect analysis

Although this research has made some progress in solving the common problems, there are some problems which are difficult to solve, especially the problem of the user's privacy protection, security and data noise.

4.1 Privacy Issues

The recommended method of this study first of all users browse the Web site information mining processing, in order to obtain the user's browsing record, get the user stealth preference information, which will inevitably get the user's hidden information, how to deal with this kind of information need our website manager's special attention. Nowadays, there are many losses caused by the leakage of users ' information, how to avoid the occurrence of this kind of problem deserves our attention. Especially in the process of information mining, we need to establish the protection mechanism of user's privacy information to ensure that the user's secret information is not leaked. Therefore, in the context of user privacy issues, the user's information needs to be strictly confidential, must take protective measures.

4.2 Security Issues

Analysis of user browsing records, but also will be the user's other important information, such as user IP, user phone number, address and other sensitive information, once leaked, will be used by unscrupulous elements, causing security risks. When using this kind of recommendation system, the website should strictly manage the user information and prevent the user from leaking information.

4.3 Noise problems with data

The characteristics of the Internet itself determine that everyone can make and deliver information on the Internet, and the publisher of the information sometimes cannot determine the correctness of the information, which accompanies a large number of erroneous data and information. Because this clustering uses the user behavior log recessive data, when obtains the user's behavior log, in addition to the research need main content, including also includes the advertisement, the copyright and so on information, this enhances the data collection the difficulty. In addition, the aimlessly roaming of Web users creates a lot of useless information and it is difficult to ensure that the data obtained is true and reliable.

5. Summary

With the popularization of Information technology and network commerce, the scale of network information will surely show an explosion, and personalized recommendation system can effectively realize the interaction with users, help users to find their favorite items quickly and accurately when browsing the website. For clustering algorithm, the target data are calculated from the page content that the target user viewed and the user's browsing path on the Internet, and the data indexes of the two aspects can be supplemented to get more accurate results. According to the content of browsing page and the weight of browsing path, the similarity between users is calculated, and the improved K-means algorithm is applied in clustering to achieve better clustering effect.

The proposed algorithm of user clustering in this paper can effectively solve the problems of the traditional recommendation algorithm in large amount of data, and not only improve the computational dimension, but also the correctness of the recommendation. The use of user behavior information data has improved the use of explicit data, and has a positive effect on large-scale promotion of recommended algorithms.

Acknowledgments

This work is supported by the Natural Science Foundation of China (nos.61472231, 61502283, 61170038), Ministry of Education of Humanities and Social Science Research Project, China (12YJA630152), Social Science Fund Project of Shandong Province, China(16BGLJ06, 11CGLJ22).

References

[1] Oldale A, Oldale J, Reenen J V, et al. COLLABORATIVE FILTERING: US, WO/2002/ 010954[P]. 2002.
[2] Ekstrand M D, Riedl J T, Konstan J A. Collaborative Filtering Recommender Systems[J]. Acm Transactions on Information Systems, 2004, 22(1):5-53.
[3] Hartigan J A, Wong M A. Algorithm AS 136: A K-Means Clustering Algorithm[J]. Journal of the Royal Statistical Society, 1979, 28(1):100-108.
[4] Tversky A. Features of similarity.[J]. Readings in Cognitive Science, 1988, 84(4):290-302.
[5] Witten, Frank I H. Data Mining[J]. Practical Machine Learning Tools & Techniques with Java Implementations, 2005, 13(4):1-1.