The Principle and Algorithm Realization of Electronic Commerce Recommendation System

Wei deng-feng[*]

College of computer science, Yangtze University, No.1 Nanhuan Road, Jingzhou, China
*Corresponding author: weidengfeng@126.com

Abstract This paper adopts the hot list and label system to solve the cold start problem, for the user reaching the product page through search engine, its recommendation is based on the solr full-text IV search engine and the search keywords related to the products. Mining user behavior pattern by mixed mining way, and implicit and explicit information to determine the user preferences, and the construct of user behavior is based on the vector space model. To calculate the keyword weight of product features, this paper use HTTPCWS system of Chinese word segmentation. Under the analysis of the advantages and disadvantages of each recommendation algorithm, combining with the actual data, this article presents combined weighted algorithm to achieve personalized recommendation; In the end this paper realize a personalized product recommendation system based on the platform of People Mall e-commerce.

Keywords: e-commerce, personalized recommendation system, user behavior patterns

1. Introduction

The task of the recommendation system is to solve the problem that the search engine's screening effect is not good when users can not accurately describe their needs. Contact with users and information, on the one hand to help users find valuable information on their own, on the other hand, so that information can be displayed in the crowd interested in him [1], so as to achieve a win-win information providers and users. The recommendation system can be divided into data layer, trigger layer, fusion filter layer and ranking layer. Data layer is composed of data generation and data storage, mainly using various data processing tools for cleaning the original log, formatted data, fall into different types of storage systems, for the use of the algorithm and model of downstream. Candidate set trigger layer is mainly used to generate the recommended candidate set from the user's historical behavior, real time behavior and geographical position. Candidate set fusion and filtering layer has two functions, one is on the starting layer gives rise to a different candidate sets were fused, recommended to improve the coverage and accuracy of strategy [2]; also any certain filtering functions, from the products, the angle of the operation of the identified some manual rules to filter out does not conform to the conditions of the item. Scheduling layer mainly uses machine learning model to re rank the candidate set

This paper does not need to consider the data generated, it is possible to consider the storage, temporarily do not consider. So the whole process is first to analyze the data, then the data pretreatment, in the candidate set trigger link, consider using collaborative filtering and position clustering method recommended set, and then by machine learning training method to obtain the final results.

2. Theoretical Analysis

Active users behavior data recorded users in e-commerce platform based on different aspects of the behavior. These behaviors for the candidate set trigger algorithm in off-line calculation (mainly browse, order), on the other hand, these act on behalf of the strength of the intention is different, so in training reordering model can according to the different behavior set different regression target value, to describe the user behavior degree more. In addition, these behaviors of the user can also be used as the cross feature of the re ranking model, which is used for the off-line training and on-line prediction of the deal model. Negative feedback data reflects the current results may in some can not meet the user's needs, so in subsequent candidate trigger need to be considered in the process of specific factors is filtered or drop right, reduce negative factors appear again chance, improve the user experience. At the same time in the reranking model training, the negative feedback data can be as rare negative examples in the training model, these negative examples to than those after the show did not click, not order samples significantly more. User portrait is to describe the basic data of user attributes, some of which are direct access to the original data [3], some are after mining the secondary processing of data. These attributes can be for candidate trigger of the deal are weighted or drop right, on the other hand can be as the characteristics of user dimensions reordering mode.

Through the data mining can extract a number of keywords, and then use these keywords to the user to label, for the user's personalized display.

3. Recommendation Engine

3.1. Recommendation Engine is Not Recommended for Different Users of Different Data

According to the public's recommendation engine, for each user are given the same recommendation, these recommendations can be static by system administrator set manually, or all users of the system of feedback to calculate the current based on popular items [5].

Personalized recommendation engine, to different users, according to their tastes and preferences are given more precise recommendation. Then, the system needs to know the characteristics of content and user recommendation, or based on a social network, by finding the user with the same user preferences, to implement the recommendation.

This is a most basic recommendation engine classification, in fact, most people discuss the recommendation engine are will be personalized recommendation engine, because fundamentally speaking, only the personalized recommendation engine is more intelligent information discovery process.
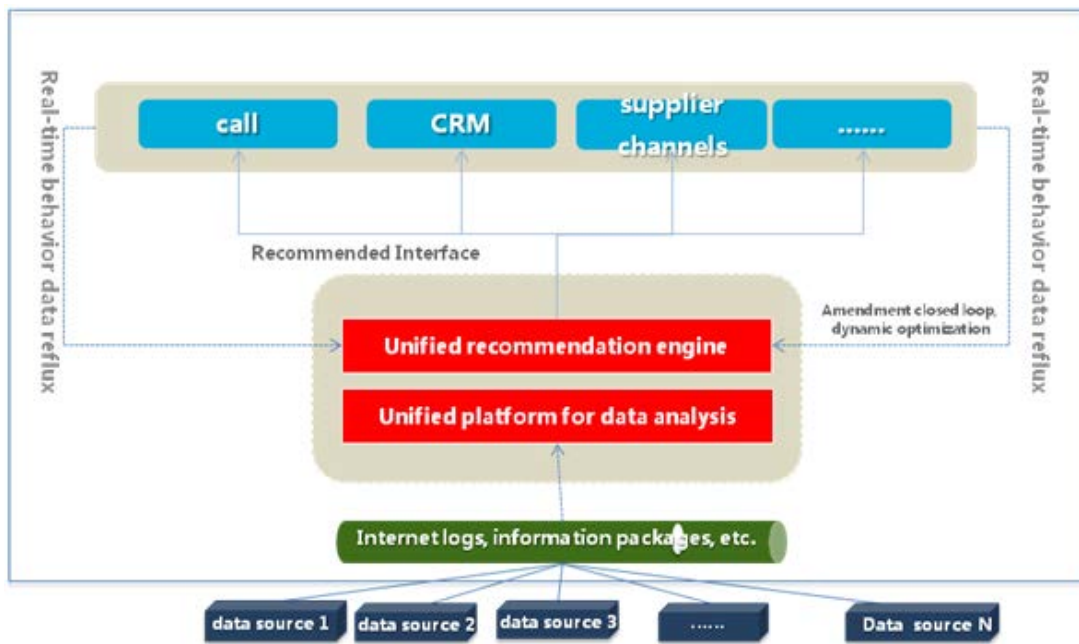

Figure 1. Recommendation system architecture

3.2. According to the Data Source of the Recommendation Engine

In fact, here is talking about how to find the relevance of the data, because most of the work principle of the recommendation engine or based on items or users of the similar set of recommendations. According to the different data sources found data correlation methods can be classified into the following categories: according to the basic information for the users of the system to find the user related degree, this is called a demographic recommendation

According to recommended items or content metadata. It is found that the goods or the content of the correlation, this is known as based on recommended content (content based recommendation) according to the user's preference for articles or information found correlations of goods or the content itself, or is the relevance of the user, this is called based on collaborative filtering recommended.

3.3. According to the Establishment of the Recommended Model

Can imagine in massive objects and the user's system, recommendation engine computation is quite large, to achieve real-time recommendation must establish a recommendation model, on recommendation model established can be classified into the following categories: Based on items and users, the recommendation engine will each user and each item as a separate entity to predict each user for preference to each item, this information is often described by a two-dimensional matrix. Since the number of users interested in the item is much smaller than the total number of items, such a model leads to a large number of data vacant, i.e., we get the two-dimensional matrix which is often a large sparse matrix. Also in order to reduce the amount of computation, we can cluster of items and users, then the record and calculation of a class of users of a class of goods preference, but such a model will loss in the accuracy of the recommendation [6].

Based on association rules recommendation (Rule-based Recommendation): association rules mining is one of the classical data mining problem, mainly mining data dependencies, the typical scenario is "shopping basket", through the mining of association rules, and I can find what items are often purchased at the same time, or the user to buy the items usually buy what other items, when we dig out the association rules, we can based on these rules recommendations to the user.

Based on model (Model-based Recommendation),: This is a typical machine learning problems, the existing user

preference information as the training samples to train a predictive user preferences model, after that the user entered the system in the can based on this model to calculate the recommended. The problem of this method is how to give the user real-time or recent preference information to the training model, so as to improve the accuracy of the recommendation.

Actually in the present recommendation system, there is little use only a recommended strategy recommendation engine, usually under different scenarios using different recommendation strategies so as to achieve the best effect of recommendations, such as Amazon's recommendation, it will based on user history itself buy number according to the recommendation, and based on the recommendation of the current user browsing items, and based on the current public preferences more popular items are in the different area recommended to the user, allowing users to can from a full range of recommended to find their own items are really interested in.

4. Algorithm Principle

4.1. Recommendation Based on Population Statistics

Is a very easy to implement the recommendation method based on demographic recommendation mechanism Demographic-based recommendation, it is simple according to the basic information for the users of the system to find the user related degree, then will be similar to the user's favorite other recommend items to the current user. The Figure 2 shows the recommended by the principle:
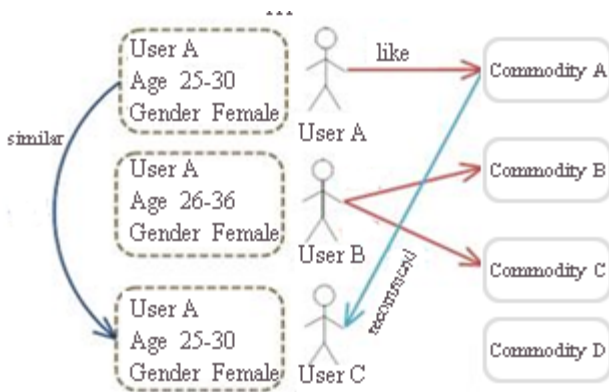

Figure 2. Working principle diagram of recommendation mechanism based on population statistics

You can clearly see from Figure 2, first of all, the system of each user have a user profile modeling, which includes the user's basic information, such as the user's age, gender, and so on. Then, the system will similarity computing the user according to the user's profile, you can see a user profile and user C, then the system will consider user a and C are similar to the user, in the recommendation engine, can say they are "neighbors". Finally, "neighbors" user group preferences to recommend to the current user objects based on, graph user will be a favorite item a recommended to the user C.

So what are the drawbacks and problems of this approach? This method based on the user's basic information to classify users is too rough, especially for the areas of high taste requirements, such as books, movies and music and other fields, can not get a good effect. May be in a number of e-commerce sites, this method can give some simple recommendations. Another limitation is that the method may involve some information that is not related to the information itself but more sensitive information, such as the age of the user, the user information is not very good.

4.2. Content Based Recommendation

Content based recommendation is the most widely used recommendation mechanism in the recommendation engine appeared at the beginning of the application, its core idea is according to the recommendation of goods or content metadata. It is found that the goods or the content of the correlation, then based on the user's preferences records in the past, it is recommended to users with similar items. Figure 3 gives the basic principles of content-based recommendation.
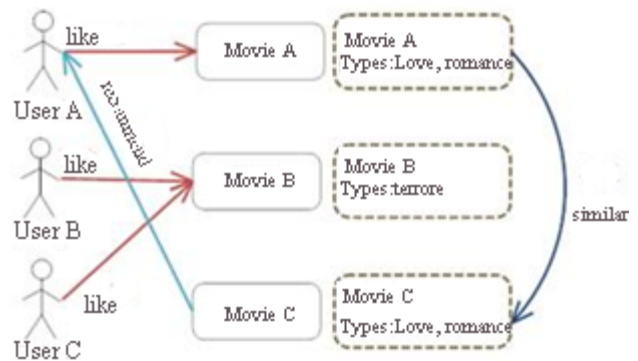

Figure 3. The basic principle of content recommendation mechanism

Figure 3 shows based on the recommendation of a typical example, a movie recommender system, first of all, we need the metadata of the film a modeling, here only simple gave a description of the type of movie; and then through the movie metadata found similarity between film, because the type is "love, romantic" film a and C is considered to be similar to the movie (of course, only the type is not enough to get a better recommendation, we can also consider the film director, actor, etc.); finally to implement the recommendation, to a subscriber. He likes movies a, then the system can recommend similar C movies to him.

The advantage of this content based recommendation mechanism is that it can provide a more accurate model of the user's taste, and can provide more accurate recommendations. But it also has the following problems:

1. Need to analyze and model the items, the quality of the recommendation depends on the integrity and overall degree of the model. In the present application, we can observe the key words and tags (Tag) is considered as a simple and effective method to describe the metadata of the objects.

2. The analysis of the similarity of objects is only dependent on the characteristics of the article itself, there is no consideration of the attitude of the person to the item.

3. Because the user needs to make recommendations based on the history of the past, so for the new user has a cold start problem.

Although the methods have many shortcomings and problems, but he was successful in a number of films, music, books of social networking site, some sites please professional personnel gene encoding of the items, such as Pandora, said in a report that in Pandora's recommendation engine, each song has more than 100 metadata features, including the song style, year, singer and so on.

4.3. Recommendation Based on Collaborative Filtering

With the development of Web, Web3.0 site more to promote user participation and user contributions, and therefore based on collaborative filtering recommendation mechanism for transportation and health. Its principle is very simple, is according to the user's preference for articles or information found correlations of goods or the content itself, or correlation is found in the user's, and then based on these relations are recommended. Based on collaborative filtering recommendation can be classified as three sub categories: one is User-based Recommendation and item based recommendation and model-based recommendation. Here we have a detailed introduction of the three collaborative filtering recommendation mechanism.

4.4. Collaborative Filtering Recommendation Based on User

Based on the basic principle of user collaborative filtering recommendation is that according to the preference of the goods and / or information to all users, and "neighbors" similar to current user tastes and preferences of users, in general use are calculated by using the k-nearest neighbor algorithm; then, based on the k a neighbor's historical preferences, for the current user recommended. Based on user collaborative filtering

recommendation mechanism basic principle, if user a like item a items C, user B like item B, C users like item a, item C, and items D; from the user's preferences in history information, we can find the tastes and preferences of users a and C is compared with similar, and C users like items D, then we can infer a user may also like items D, so you can d items will be recommended to a user. Users of collaborative filtering recommendation mechanism and recommendation mechanism based on demographic is calculate user similarity based on and based on "neighbors" user group recommended by calculation, but they are different is how to calculate user similarity, mechanism based on demographic only consider itself characteristics, and user based collaborative filtering mechanism but on the preferences of the user's historical data to calculate the user similarity, is its basic assumptions, like items of similar users may have the same or similar tastes and preferences.

5. Concrete Model

5.1. Build Matrix

The user's score is divided into explicit and implicit ratings, this time the data only implicit rating that, browsing, collection, plus shopping cart and purchase.

Table 1. User behavior definition table

Behavior name	behavior description
view times	Value for browsing
Collection	value is (0,1), 1 for collection
Join a shopping cart	value is (0,1), 1 for join
Purchase	value is (0,1), 1 for purchase

Data processing to get structured data

Table 2. Structured data sheet for accessing behavior

Serial number	User id	Product id	Browsing times	Collection	cart	buy
1	User 1	Item 1	1	1	1	0
2	User 1	Item 2	5	1	2	1
3	User 2	Item 1	1	0	1	1
N	User N	Item 1	0	0	0	1

Assume that M represents the number of users and N represents the number of goods; on behalf of the account for goods j the actual score, 1 and less than or equal to I is less than or equal to m, 1 and less than or equal to j is less than or equal to N; will user behavior transformation for the implicit rating rules are as follows:

1) If the user i to buy a commodity J, it is equal to 5;

2) If the user i added j to the shopping cart, it is equal to 4;

3) If the user i will add J commodity favorites, it is equal to 3;

4) If the user J on the commodity I browsing for more than 2 times, it is equal to 2; if the number of clicks for the 1 time, it is equal to 1;

Scoring rules can be adjusted by the accuracy of the recommended results; Usually the user for a commodity will do a number of operations at the same time, such as a user to click on goods, adds the favorites, then add the shopping cart, and eventually purchased, which scores the

highest. Then we can establish the user - commodity score matrix

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & \ddots & \vdots \\ y_{m1} & \cdots & y_{mn} \end{pmatrix}.$$

5.2. Time Dimension

5.2.1. H.Ebbinghaus Forgetting Curve

The user's interest is dynamic change, the user's recent visit and the score of goods more able to reflect the current interests of users, but also affect the user's current purchase decisions. The early access of goods for the user's current interest may have a small impact on the impact of the role of goods, that is, the user's access behavior and the importance of the rating will continue to decay over time. Consumer behavior can be considered as

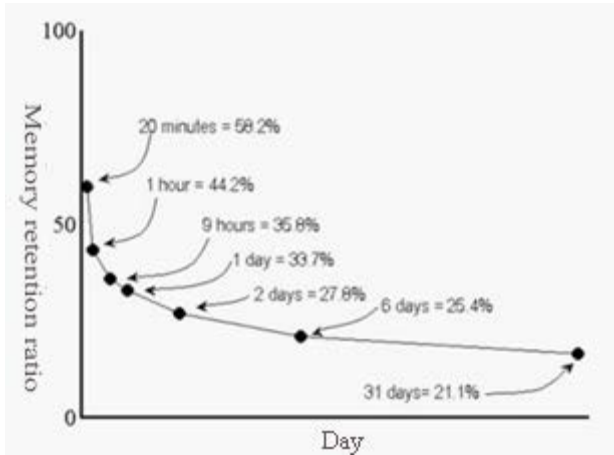a kind of psychological behavior, according to the law of forgetting curve.

Figure 4. Forgetting Curve

Represents the start time, which indicates the time for the project. The score time is the user's comprehensive evaluation of the goods time, that is, the time of the occurrence of behavior. T indicates the interval between the user's rating time and the effective start time.

The exponential function formula (1) which indicates the change of user's interest is shown below:

$$f(t) = e^{-\lambda t}. \tag{1}$$

In the formula, the weight (0, 1), can be adjusted according to the accuracy of the results. The greater the more, the interest of the time decay faster, and vice versa.

5.2.2. Recommended Process

Step 1: using the improved Pearson correlation coefficient formula to calculate the similarity between the two users, the formula is as follows:

$$sim(a,b)$$

$$= \frac{\sum\limits_{j \in Iab} \left(y_{aj} \times f(t) - \overline{y_a}\right)\left(y_{bj} \times f(t) - \overline{y_b}\right)}{\sqrt{\sum\limits_{j \in Iab} \left(y_{aj} \times f(t) - \overline{y_a}\right)^2} \sqrt{\sum\limits_{j \in Iab} \left(y_{bj} \times f(t) - \overline{y_b}\right)^2}}$$

y_{aj}, y_{bJ} respectively score of user a and user B of commodity J, Iab said user a and user B common rated items set, f (T) as a forgetting function, user a score of

commodity aggregate average score, user B score commodity aggregate average score.

Step 2: the first k users with the highest similarity of the user U as its nearest neighbor collection a.

Step 3: comprehensive neighbor users to evaluate the product J and predict the user a on the product J score. Assuming that C represents a neighbor user, PS (a, J) represents the target user's predictive score, and the formula for the prediction score is as follows:

$$PS(a,j) = \frac{\sum\limits_{c \in U} \left(p_{cj} - \overline{p_c}\right) sim(a,b)}{\sum\limits_{c \in U} sim(a,b)} + \overline{p_a}$$

Step 4: The highest prediction score the first n commodities as automatic goods..

6. Conclusion

Based on data, using algorithms to carve, only to combine the two, will bring to enhance the effect. For us, these two nodes are optimized our milestone in the process: These are summarized in us a little practice, of course, we are still on the way.

References

[1] D. Jannach, M. Zanker, A. Felfernig, G. Friedrich. Recommender Systems: An Introduction [M].Cambridge University Press. 2011.

[2] R. Lambiotte, M. Ausloos, Collaborative tagging as a tripartite network, Arxiv preprint cs. DS/05 12090, 2005.

[3] S. Golder, B.A. Huberman, The Structure of Collaborative Tagging Systems, Arxiv preprint cs. DL/0508082, 2005.

[4] Elahi, Mehdi; Ricci, Francesco; Rubens, Neil. A survey of active learning in collaborative filtering recommender systems. Computer Science Review, 2016, Elsevier.

[5] Beel, Joeran, Langer, Stefan, Genzmehr, Marcel (September 2013). "Sponsored vs. Organic (Research Paper) Recommendations and the Impact of Labeling". In Trond Aalberg; Milena Dobreva; Christos Papatheodorou; Giannis Tsakonas; Charles Farrugia. Proceedings of the 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013) (PDF). pp. 395-399. Retrieved 2 December 2013.

[6] Tuukka Ruotsalo; Krister Haav; Antony Stoyanov; Sylvain Roche; Elena Fani; Romina Deliai; Eetu Mäkelä; Tomi Kauppinen; Eero Hyvönen (2013). "SMARTMUSEUM: A Mobile Recommender System for the Web of Data". Web Semantics: Science, Services and Agents on the World Wide Web (Elsevier) 20: 657-662.