SHS Web of Conferences **14**, 01002 (2015)
DOI: 10.1051/shsconf/20151401002

Research on Algorithm Recommended by Online Education for Big Data

Tao Feng and Yun Cheng

Liaoning Economic Vocational Technological Institute, 110122 Shenyang Liaoning, China

Abstract. "Big data" is becoming a hot topic in the Internet. The long tail problem of the massive online courses also becomes the biggest headache for operation team of online education. The manner in which the reader wants most courses show to be presented before the user is the key to improve the quality of online education. Personalized recommendation system is to discover the readers interests tendency based on the existing user data, project data, and interactive data, thus to provide personalized product recommendation for readers. This article is based on the two kinds of algorithms, namely the content and the collaborative filtering recommendation to propose an improved integration scheme, which can make good use of existing data to discover the useful knowledge for readers' recommendation. The method firstly solves the sparsity problem in traditional collaborative filtering, and meanwhile we start from the global structure relation of course, to analyze the relationship between the reader and the course more comprehensively. The algorithm to improve the accuracy of recommendation from multiple angles, and provides a feasible method for precise recommendation of online educational video.

Keywords. recommendation algorithm; user interaction; online education; collaborative filtering recommendation; content recommendation

1 Introduction

Information retrieval and information recommendation is the main tool to solve the problem of big data[1]. Information retrieval is to weed out irrelevant information by providing keywords, and then dig out the relevant content from a mass of information. This method is more suitable for user with clear purpose. While for users with uncertain demand, they only want the system to be in accordance with their own interest or historical operation records to recommend some information that may be interesting for them, and thus to provide a better user experience and higher work efficiency. Personalized recommendation system uses personalized recommendation algorithm to make an analysis and research on content feature information, score or history operation records and other information of users and items, and discover the interests of users to screen the item set and provided personalized recommendation for specific users, thus to solve the problem of information overload[2].

This article analyzes the current technical tool used for personalized recommendation engine, constructs a universal personalized recommendation system of RC version. We introduce Hadoop, Storm, RabbitMQ, Redis technology and proposed a complete design schedule for recommendation system, in which the ideas of realization respectively for the off-line processing of big data and real-time computing are put forward. In addition, the module communication message based on message queue is introduced in communication between modules, which ensures the processing capacity and real-time performance for system.

2 Personalized data and its analysis

In order to provide effective and accurate recommendation set to users, while guaranteeing the performance of the recommender system and other non-functional requirements, the researchers and enterprises have put forward many personalized recommendation algorithms, such as Item-based collaborative filtering recommendation algorithm[3,4], User-based collaborative filtering recommendation algorithm[5], Content-based recommendation algorithm[6,7,8], Cluster-based collaborative filtering recommendation algorithm[9], SVD-based collaborative filtering recommendation algorithm[10] and image-based collaborative filtering recommendation algorithm[11-12], etc.. These algorithms uses data mining techniques to conduct in-depth analysis of user data and project data to obtain the interest characteristics and the specific patterns of behavior for users, and thus to provide personalized recommendation for users. The personalized recommendation algorithm based on data mining consists of two stages of learning and use. In the process of learning, personalized recommendation algorithm conducts mining

analysis on the original data, and establishes the recommendation model corresponding to algorithm. The recommended model data can be used to provide personalized recommendation for real-time guidance of users in the stage of use.

Recommendation algorithm based on the content is derived from the traditional information search technology, and it depends on the system to extract the project features, analyze the user behavior, and research the Internet users' interests and preferences to provide item set with similar features to them. The algorithm does not rely on the historical data of score between the user and the project[5].

Personalized recommendation algorithm based on collaborative filtering is the most valuable recommendation algorithm in the field of research and enterprise application fields at present. Personalization is its main goal to be realized. As for the difference from the classic content-based recommendation methods, the algorithm is mainly to conduct analysis and mining the user groups with high similarity to the target users, or item set similar to target item, and then use the user group and item set to provide personalized recommendation for users. According to the difference of used business association, the collaborative filtering recommendation algorithm can be divided into User-based collaborative filtering algorithm [5], Item-based collaborative filtering algorithm[8,4] and Model-based collaborative filtering algorithm[11-12], etc..

Each algorism in personalized recommendation system has its advantages and disadvantages, and also a certain degree of complementarity in preferences. So In the current Web recommendations will not adopt one single recommendation mechanism and strategy, but to integrate multiple methods, namely Hybrid Recommendation, thus to achieve a better effect of recommendation [13,14,15]. There are many combinations of hybrid recommendation, and the specific combination principle will be varied with different data and scenes. Therefore, we should choose the right combination of methods to achieve the full effect.

3 Integrated personality recommendation algorithm

This article improves the hybrid recommendation approach. Based on bipartite graph, we first use the user's history score information and item category feature information to construct a graph model based on user and item; the random walk algorithm will be used for computing global similarity between items in the graph model. This method has low computational complexity.

3.1. Two-layer weighted graph model

Let G={V, E, W} as a weighted mixed graph, among which $V = V_{user} \cup V_{item}$ represents the vertex set, V_{user} the user vertex, V_{item} the item vertex; $E = E_H \cup E_{UI}$ the edge set, E_H the inner edge of item, E_{UI} the connecting edge for user and item; and W the edge weights set. At this time, the mixed weighted graph can represent a two layer model, in which the upper part is user layer, and the lower part item layer, as shown in Figure 1.

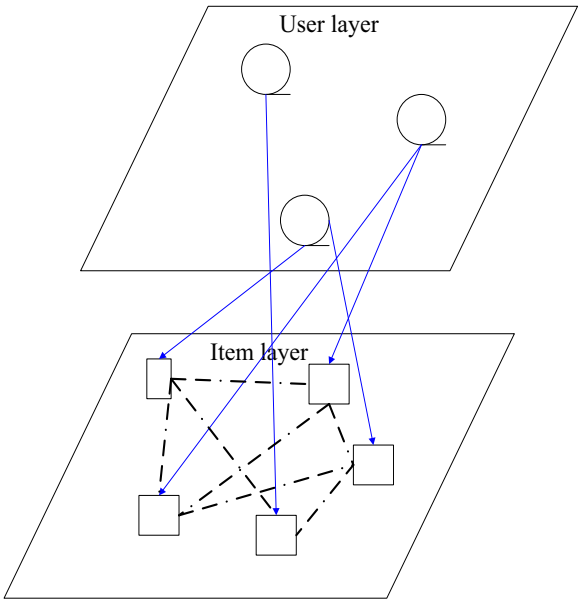

Figure 1. Two-layer model of user-item.

The edge set E_{UI} is the connection between User layer and Item Layer. At this time, the set W_{UI} is the degree set of that connection. Let w_{ui} as the connection degree of User u and Item i, which represents the user preference for the item or the size of score, namely $w_{ui} = r_{ui}$. Higher score will lead to higher degree of edge. However, in the electronic commerce system in score class, the trend of information exist in the original score information, so it cannot accurately represent user preference for the item. Take the reference points of user ratings as an example: reference points for part of the user score are higher, for example, 3 points are the reference point for Like, 2 points for Dislike. Similarly, some movie items, compared to others, tend to have higher score, and it may be affected by the release time of movies. This kind of trend information can be called the Global Effect (hereinafter referred to as GE)[16]. Before the score is taken as a preference degree of users for items, first of all we need to remove the global effect from the original score. In the actual recommendation system, there are so many influence elements of these global effect, such as holiday, quarter, etc. that affect the score over the same period. But the experiment proved that the effect of reference points GE on the final score is the biggest. This article only considers the global effect for two kinds of reference points: reference point for user and item. Let the reference point of User u and Item i respectively as GE_u and GE_i, and they can be represented in a manner shown in formula (1):

$$GE_u \frac{\sum_{\{u,j\}\in\rho} r_{ur}}{k(u)}, GE_r \frac{\sum_{\{u,j\}\in\rho} r_{ur}}{k(i)} \quad (1)$$

Among which, ρ represents the score set, k(u) the scoring times of User u, k(i) the scoring times of Item i. By removing those two reference points GEs, we can get the preference degree of User u and Item i $w_{ui} = r_{ui} - GE_u - GE_i$, and use linear normalization method to implement normalization processing for w_{ui}, as shown in formula (2).

$$w_{ui} = \frac{w_{ui} - w^u_{\min}}{w^u_{\max} - w^u_{\min}} \tag{2}$$

In the formula above, $w^u_{\max}$ and $w^u_{\min}$ respectively represent the maximum and minimum score of user after removing reference points.

The edge set E_H is the internal connection of Item Layer. At this time, the set W_H is the degree set of that connection. For those two items, the degree between them can be represented as a similarity degree between any two nodes. The higher similarity will lead to greater degree. At this time, assuming that w_{ij} represents a connection degree between Item i and j, we can use the comprehensive similarity of Movie Lens system $sim_s(i, j)$ to represent the degree between them, namely $w_{ij} = sim_s(i, j)$.

3.2 Random walk and recommendation algorithm based on weighted two-layer graph

In physics, random walk is presented as a kind of irregular forms of motion. Each step of its motion is random and independent from the previous transfer[17]. In the two layer graph model of user-item, the relationship between the user and the item is represented as a random process $\{X_n\}$, and its state space is the node in the two layer graph. If the current node is i (possibly is the node of user or item), the walk between node i and its connection node is random. Let the next walking connection node as j, walking probability $P_{i,j}$, while the degree of this probability is only related to the nearest m walking nodes. We call the walking trace $\{X_n\}$ as Markoff chain. If m=1, the next state value X_{n+1} is only related to the current state value X_n, as shown in formula (3).

$$P_{i,j} = P\{x_{n+1} = j \mid x_n = i\} \tag{3}$$

This $\{X_n\}$ is the first-order Markoff chain, referred to Markoff chain. $P_{i,j}$ represents one-step transition probability from node i to node j. Markoff chain is to predict the next step of the nodes according to the current nodes and the current one step transition probability. And random walk model is a classical balanced Markoff chain, so the random walk algorithm can finally get a balanced state[17-19]. This article uses the random walk algorithm, starting from the user node, to get the similarity between the user node and all other nodes. Finally, we can use the size of the similarity to select 2 users or items with largest similarity and relevant to the users, namely the Top-K user and item recommendations.

In the random walk model of the mixed weighted graph G, we can use the weight w_{ij} between two points in the graph to indicate the direct transfer probability P_{ij} between two nodes, as shown in formula (4).

$$P_{ij} = \frac{w_{ij}}{\sum_{k=1}^{n} w_{ik}} \tag{4}$$

$P_{ij}(k)$ represents the probability of random walking from node i to node j through node k, namely two-step transfer probability. It can represent the product of probability of direct transfer from node i to node k P_{ik} and the probability of direct transfer from node k to node j P_{kj}; Similarly, we can use $P^t_{i\to j}$ to represent the probability of transfer from node i to node j by t steps, namely $P^t_{i\to j} = P^{t-1}_{i\to k} \times P_{kj}$.

By t steps of random walk, the similarity between any two nodes will reach a stable equilibrium, and at this time, we can use the probability of random walk from node i to node j to represent their global similarity, namely $RD(i, j) = \sum_{k=1}^{t} \delta^k P^k_{i\to j}$, among which δ is used to balance the weight of transfer probability for each step, The multiple-step transition probability can go through weighted combination by the parameter δ; t represents the step of random walk, and higher value of t indicates that there are more neighbour information.

If node i and node j are both item nodes, RD(i,j) will be the global similarity between those two items. At this time, we can use RD(i,j) to conduct Top-K recommendation based on item similarity for users. If two nodes were user node u and item node j, RD(i,j) will be the global similarity between user and this item. At this time, we can directly use this value to conduct Top-K recommendation for users.

4 Algorithm implementation and experiment

4.1 Algorithm implementation

Algorithm implementation, in physical structure, is composed of a Hadoop cluster, Storm cluster, Redis cache cluster, message queue middleware, Web service cluster and multiple database clusters. In a real environment, the cluster size has great influence on the bearing capacity and the processing ability of the system. This article uses a single or several servers to simulate the work of cluster. Table 1 provides the servers and the system where they run, in the process of algorithm implementation.

Table 1. Servers and the system where they run, in the process of algorithm implementation.

The host IP and port	Operated process	Operating system	Function of host system
192.168.1.71:6080	Redis	Centos	The front-end cache cluster
192.168.1.71:6081	Redis		
192.168.1.71:6082	Redis		
192.168.1.72:5670	RabbitMQ	Fedora	Message queue middleware
192.168.1.72:6080	Redis		
192.168.1.70:80	Apache	Centos	WEB service
192.168.1.70:3306	Mysql		Database
192.168.1.73	Hadoop-NameNode Storm-Nimbus	Cnetos	Hadoop Name Node Storm Nimbus node
192.168.1.74	Hadoop-DataNode Storm-Supervisor	Centos	Hadoop data node Storm Supervisor node
192.168.1.75			
192.168.1.76			

4.2 Experimental data

This article uses the experimental data set provided by the online education platform, including basic information of users up to 50000, basic information of 3000 films and the scoring information of 10000 users on the film, among which the score value is an integer within an interval of[1, 5], and higher value indicates a higher preference degree of users for the items.

In recommendation system, the measurement methods of recommendation quality of recommendation algorithms include measurement method of decision support accuracy and the measurement method of statistical precision and other methods. Common metrics include accuracy, novelty, etc[20]. This article uses the mean absolute error (MAE) to measure accuracy.

4.3 Experimental results and analysis

1. Calculation of mixed similarity

Item similarity is composed of potential similarity and similarity based on item characteristics. By introducing the parameter β, the potential similarity $sim_a(i,j)$ and feature similarity $sim_c(i,j)$ can be combined, in order to obtain the hybrid similarity between items. This experiment is performed under the condition of α=1900, and calculates the recommendation accuracy through the optimal value selected by iterative parameter. The experimental result is as shown in Figure 2.

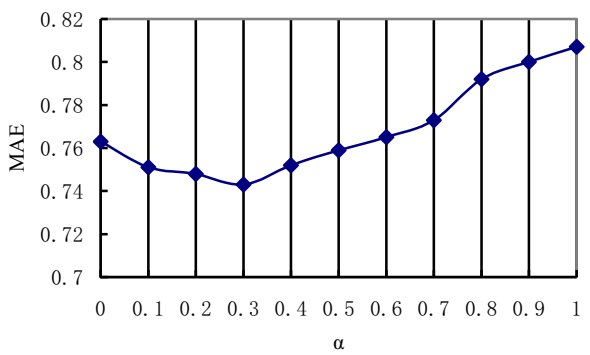

Figure 2. Experimental results of item similarity.

As shown in Figure 2, when the parameter β=0, namely it just uses the potential similarity as the final similarity, at this time MAE≈0.763. When the parameter is incremented and MAE is significantly reduced accordingly, it will prove that this mode of mix is effective. However, when the parameter β exceeds a certain value, MAE will be conversely increased. When the parameter β=1, namely it only uses the similarity based on item types to measure the similarity among items, at this time MAE≈0.807. When β=0.3, MAE≈0.743 will be the minimum value, namely the optimal mix.

2. Parameter selection of random walk for each weight

In the random walk algorithm, each step of the walk represents a certain degree of similarity between two items. At this time, the similarity generated by each step can be weighted using the parameter δ, and use this weight to conduct weighted combination for the values of all steps, thus to get the final calculation result. This experiment constructs two-layer model by fixed parameters α=1900 and β=0.3, and the step is selected as t=5. The experimental result is as shown in Figure 3.

As shown in the experimental result, when 0<δ<0.6, MAE will be reduced with the increase of δ value. When δ>0.6, MAE will be conversely increased with the increase of δ value. When δ=0.6, MAE≈0.715, and is also in its minimum value.

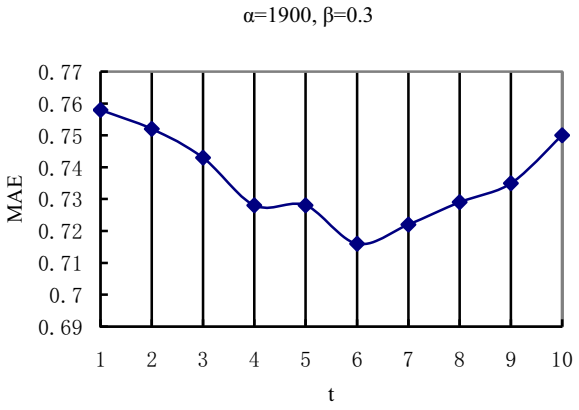

Figure 3. The experimental results of parameter selection for each step of weight.

5 Conclusion

This article presents a new hybrid recommendation model based on the advantages and disadvantages of recommendation method of content and collaborative filtering. The model can make full use of rating information and feature information between users and items, and consider the user-item relevance based on the relation of global structure. Compared with the traditional algorithm, the hybrid recommendation model is improved to a certain extent in the recommendation accuracy and recommendation efficiency method, but it is still not comprehensive. First of all, subject to the experimental data, this article only considers the characteristic of the item category, but in the real electronic commerce system, the relationship between the items is perplexing, and we can't just use classes to well measure the relationship between items. The next step, we can introduce the cognitive computing and other methods that are close to the human thinking to comprehensively consider the relation between items.

References

1. C.D. Manning, P. Raghavan & H. Schutze. *Introduction to Information Retrieva* [M]. England: Cambridge University Press, 2008.
2. J.B. Schafe, L.J. Konstan & J. Ried. E-commerce recommendation applications[J]. *Data Mining and Knowledge Discovery*, 2001, 5(1-2):115-153.
3. B.M. Sarwar & Karypis. Item-based collaborative filtering recommendation algorithms [C]. *Proceedings of the 10th International Conference on World Wide Web*, N.Y, USA: ACM, 2001: 285+295.
4. G Linden, B. Smith & J. York. Amazon.com Recommendations: Item to item collaborative filtering [J]. *IEEE Internet Computing*, 2003, 7(1):76-80.
5. Zhao Chenting & Ma Chune. Exploring the Secret inside the Recommendation Engine: In-depth Study on the Algorithm Relevant to Recommendation Engine--collaborative filtering [Online]Available from: http://www.ibm.com//developerworks//cn//web//1103_zhaoct_rec ommstudy2//.
6. M. Balabanovic & Y. Shoham. Fab: Content--based, collaborative recommendation [J]. *Communications of the ACM*, 1997, 40(3):66-72.
7. R. Melville & R.J. Mooney. Contented--based collaborative filtering for improved recommendations [C]. *Proceedings of the 18th National Conference on Artificial Intelligence*, 2002, pp.189-192.
8. B.M. Kim, Q. Li & C.S. Park. A new approach for combining content--based and collaboration filters [J]. *Journal of Intelligent Information System*, 2006, 27(1):79-91.
9. H. Lyle & PI Dean. Clustering methods for collaborative filtering[C]. In: Workshop on Recommendation Systems at the Fifteenth National Conference on Artificial Systems, 1998, pp.114-129.
10. M. Vozalis & K. Margaritis. Applying SVD on Item--based filtering[C]. Proceedings of the 5th International Conference on Intelligent Systems Design and Applications, 2005, pp.464-469.
11. R. Paulson, A. Tzanavari. Combining collaborative and content--Based filtering using conceptual graphs [J]. *Modelling with Work*. 2003:168-185.
12. Huang Zan & Chung Wingyan. A Graph Model for E-Commerce Recommender Systems[J]. *J. ASIST*, 2003, 55(3):259-274.
13. R Resnick & H.R Varian. Recommender systems[J]. *Communications of the ACM*, 1997, 40(3):56-58.
14. Wu Lihua & Liu Lu. Modeling overview of personalized recommendation system users [J]. *Information Journal*, 2006, 25(1):55-56.
15. G. Adomavicius & A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state of the art and possible extensions [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2005, 7(6):734-749.
16. R.M. Bell & Y. Koren. Improved neighborhood--based collaborative filtering [C]. KDD Cup'07, San Jose, California, USA, August 12, 2007, pp.7+1 4.
17. S. Baluja & R. Seth. Video suggestion and discovery for youtube: Taking random walks through the view graph [J]. *Proceedings of 17th Intel World Wide Web Conference*, 2008, pp.895-904.
18. F. Fouss & A. Pirotte. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation [J]. *IEEE Transactions*, 19, 2007:355-369.
19. Zhou Junjun, Wang Mingwen & He Shizhu. Collaborative filtering recommendation algorithm based on random walk and clustering smoothing [J]. *Journal of Guangxi Normal University* (Natural Science Edition), 2011, 29(1):173-178.
20. Liu Jianguo, Zhou Tao, Guo Qiang & Wang Binghong. Overview of evaluation methods of the personalized recommendation system [J]. *Complex Systems and Complexity Science*, 2009, 6(3):1-10.