Research on Course Resource Recommendation System Based on Feature Selection

Hu Anming

School of Computer Science & Engineering, Tianhe College, Guangdong Polytechnic Normal University, Guangzhou, 510540, China;

Keywords: learning resources, TF-IDF algorithm, recommendation algorithm

Abstract: The TF-IDF algorithm is used to calculate the weights of various learning resource keywords. It combines the learner learning preference with vector matrix to generate the weight vectors of learning resource preference. In addition, a personalized recommendation algorithm is used to recommend learning and course resources, reduce the time for searching resources, improve the learning efficiency of learners, fulfill the differentiated needs of learners for leaning and finally enhance learning efficiency.

1. Background and Significance

In recent years, with the rapid development of network and software technology, information technology in education has made a rapid progress. A myriad of universities have established online education platforms for their students. However, the construction of online learning resources is often divided into majors and specialized subjects and the amount of information is enormous. Meanwhile, the curriculum may somehow pay no attention to the differentiation of learners at different stages. With the continuous expansion of the scale of the curriculum resource platform and the rapid increase in the number and types of courses, learners need to spend a lot of time to find suitable learning resources. The process of spending a lot of time in browsing a great deal of irrelevant information and products will undoubtedly lead to a continuous loss of learners who are embedded in information overload problems. In a bid to solve these problems, personalized recommendation systems came into being.

How to enable learners to quickly extract the required learning resources from the vast learning resources and carry out an effective learning is a very essential research. It is urgent to implement personalized education and improve the learning efficiency of learners. Therefore, this article proposes the use of personalized recommendation technology to assist learners to find appropriate learning resources based on their own characteristics. It aims to heighten the accessibility of resources and enhance learning efficiency.

2. Personalized Recommendation Technology

The personalized recommendation system is an advanced business intelligence platform based on massive data mining. It is dedicated to provide personalized information services and decision supports for learners. It has shifted from service recommendations such as life, entertainment and learning recommendations to mobile application recommendations. At present, large-scale e-commerce companies, for instance, Amazon and eBay have used various forms of recommendation systems in varying degrees. Personalized recommendation systems can effectively retain customers and further improve the service capabilities of e-commerce systems. Moreover, various Web course resource sites that provide personalized services also need the support of recommendation systems. As a result, it is imperative to implement personalized course recommendation.

There are numerous methods to implement personalized recommendation technology. In general, they can be divided into two main categories: collaborative filtering based on historical data; collaborative filtering based on models. Among them, collaborative filtering based on historical data

  DOI: 10.25236/eeaet.2020.004

includes: User-based collaborative filtering (UserCF) and Item-based collaborative filtering (ItemCF); User-based collaborative filtering mainly depends on the historical records generated between the learner and the item, such as the learner ’s Clicks, followers, purchases, etc.. Calculating and recommending based on the historical records between learners and items; Item-based collaborative filtering mainly relies on the similarity of items to make recommendations.

Model-based collaborative filtering often uses various models to perform collaborative filtering calculations, such as regression prediction-based collaborative filtering, matrix-based collaborative filtering. With comparison to historical data-based collaborative filtering, model-based collaborative filtering have better prediction accuracy, lower spatial complexity and lower computation time. It is also the main focus of this current research.

This academic paper uses TF-IDF algorithm model to extract features from various learning resources. It combines learner interest preference matrix to calculate learner interest model with learning resource feature similarity. It hopes to achieve a feature selection-based curriculum resource recommendation system. Through recommending learning and course resources, a personalized learning resource recommendation system reduces the search time for resources, enhances the learning efficiency of learners and meets the differentiated requirements of learners for learning.

3. Feature Selection-based Recommendation Technology

Based on the feature selection collaborative filtering algorithm, the TF-IDF algorithm is first used to extract the features of various learning resources. The weights of learner's preference for different courses and learning resources are counted. A similarity matrix of preference attributes is generated based on the weights over a period of time. It solves the problem of time drift. Performing TF-IDF calculation on the course and learning resources to extract the keywords in the course content. A weighted calculation of the similarity matrix between the keywords of the course content and the learner's preference attributes aims to achieve the goal of the learning recommendation system.

The implementation process is as follows:

(1) Calculation of learner preference weights

The window control method is used to classify the interests of learner. The number of clicks and visits of the learner is arranged in ascending order. When i-th click and visit of the learner exceeds the average visits of all resources, the material is added to learner i's long-term interest set L.

Suppose the number of times a learner visits a resource is, the maximum number of visits is, and the i learner's interest weight for j resource is, then the value is

$$w'_{i,j} = \frac{n'_{i,j}}{n_{max}} \tag{1}$$

When the access frequency of the learner to the j-th attribute is greater than the threshold α, we add this attribute to the short-term interest of the learner in set S. Let the sliding time window be, the number of visits to resource i from i learner, then the weight of resource j is:

$$w''_{i,j} = \frac{n''_{i,j}}{n_a} \tag{2}$$

(2) Keyword extraction of learning resources

Keyword extraction is a critical part of processing the attributes of learning resources. In this paper, TF-IDF algorithm is used to extract the keywords of curriculum and learning resources. TF-IDF algorithm is a key technique for measuring words or phrases in text information based on statistical methods. The principle is that when a word is used more often in the target text but less frequently in the corpus, then the word can be used as the keyword of the text. TF refers to the proportion of the number of times a word appears in the target text, referred to as the word frequency. TF is calculated as follows:

Find the frequency of the word i in the text j, $\sum_{i=1}^{n} f_{i,j}$ refers to the total number of word in the document, the formula is as follows,

$$TF_{i,j} = \frac{f_{i,j}}{\sum_{i=1}^{n} f_{i,j}} \tag{3}$$

IDF refers to the inverse document frequency. The IDF value indicates the universal weight of a vocabulary. It is assumed that the smaller the number of texts containing the word i in a text corpus collection, the larger the IDF_i value of the word. The better, the more likely it is to be used as a keyword. The IDF calculation formula is as follows:

$$IDF_i = \log \frac{N}{n_i + 1} \tag{4}$$

Where N is the total number of words in the text, n_i is the number of occurrences of the word i in the text, and 1 is added to avoid the situation where the denominator appears to be 0.

If the word i appears less frequently in the text N and more times appears in the text j, then the word i can be used as the keyword of the text j. The TF-IDF algorithm formula is as follows:

$$TF - IDF_i = TF_{i,j} * IDF_i \tag{5}$$

Let the learning resource set be A = $(a_1, a_2, a_3, \ldots, a_n)$, all the set of learning resource content that introduce text word is grouped as M = $(m_1, m_2, m_3, \ldots, m_n)$, where the k th introduction of learning resource is expressed as $c_k = (w_1^k, w_2^k, w_3^k, \ldots, w_n^k)$ and w_n^k represents the weight of n th word m_n in k th learning resource. The TF-IDF formula of word m_i in k th learning resource is as follow:

$$TF - IDF(_{mi,ak}) = TF_{(mi,ak)} * \log \frac{N}{n_i} \tag{6}$$

(3) Calculation of recommending learning resources

Through the previous calculations, the keywords of the learning resource content and attributes of the learner's preference can be obtained. The cosine similarity algorithm in the collaborative filtering algorithm is used to recommend the learning resources. The cosine similarity algorithm uses the cosine angle between the preference attribute vectors to calculate the recommendation weight. In this topic, the learner's preference attribute is used to calculate the similarity of the learning resources. The formula is as follows:

$$\cos(\theta) = \frac{A.B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2 * \sqrt{\sum_{i=1}^{n}(B_i)^2}}}$$

After calculation by the above formula, the similarity of learning resources is ranked from high to low according to accuracy and recommendations are made.

(4) Feature-based course resource recommendation system

According to the implementation strategy of the feature selection recommendation system above, a feature selection recommendation system model is constructed. The model consists of four modules: course resource feature extraction, learner preference calculation, data clustering, and recommendation result generation, as shown in Figure 1:

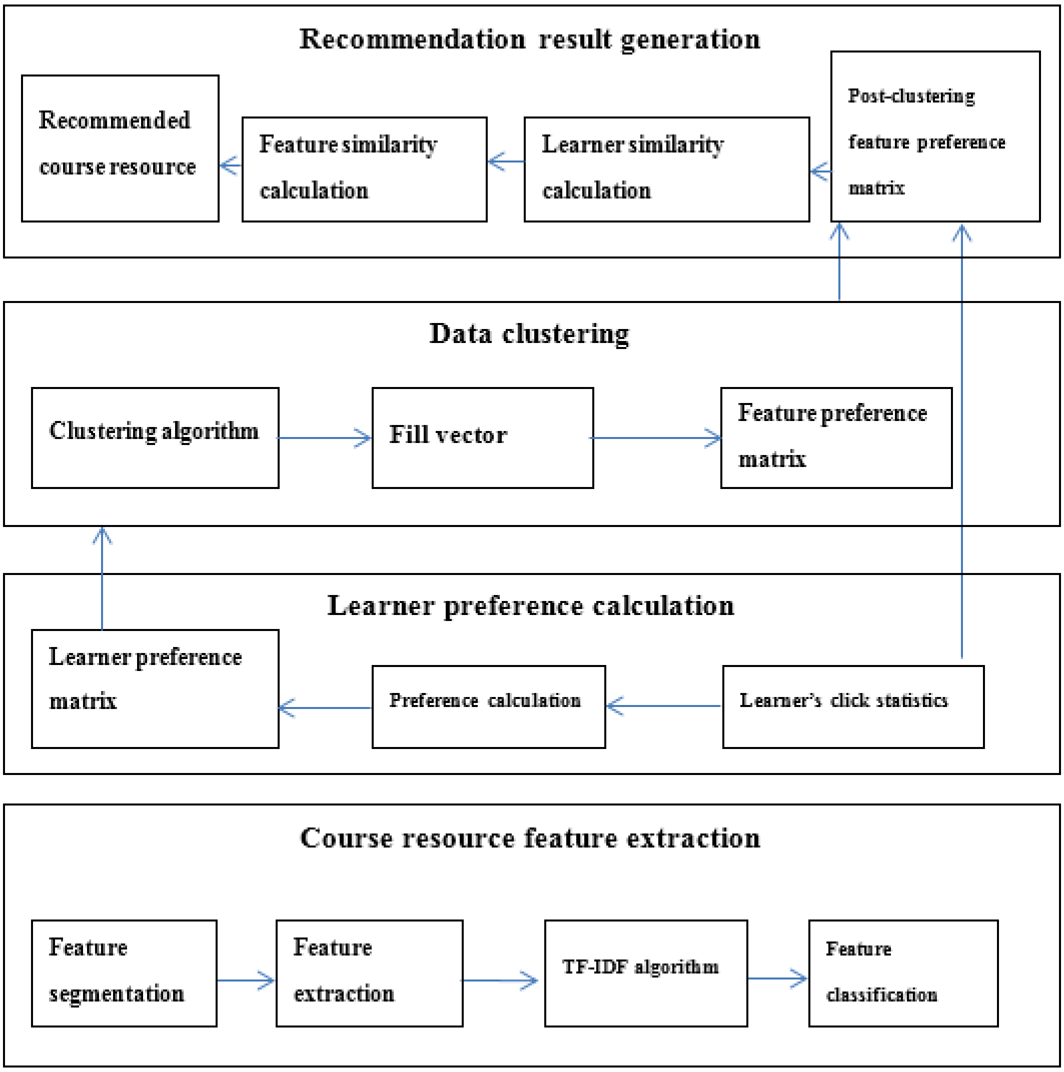

Figure 1 Recommendation result generation system

Course resource feature extraction module: Course resource feature extraction uses specific word segmentation technology to divide the course resources. News features are extracted. TF-IDF algorithm is utilized to vectorize the course resources. Moreover, a specific classification algorithm is used to classify the course resources.

Learner preference calculation module: Counting clicks of learner and arranging the click in ascending order. Then, compare the clicks of learner with the average number of visits to all resources. If it is above the average click occurrence, calculate preference weights by the information of clicks. Finally, generate a learner preference matrix.

Data cluster processing module: This module uses the K-Means clustering algorithm to cluster the interest matrix of the learner. It is a part of data preprocessing of collaborative filtering algorithm, which divides the preferences of the learner into n sub-matrixes. The learners within the same cluster will have a high degree of similarity, while the learners in different clusters have a low degree of similarity. It can reduce the time complexity of the recommendation algorithm.

Recommendation result generation: According to the learner preference matrix and the course resource feature matrix, the cosine similarity is used for generating a recommendation list in recommendation calculation.

(5) Implementation process of feature-based course resource recommendation system

The entire feature selection course resource recommendation system uses the SpringBoot framework to implement the overall framework. The Mahout package feature recommendation algorithm is also used. Its structure is shown in Figure 2:

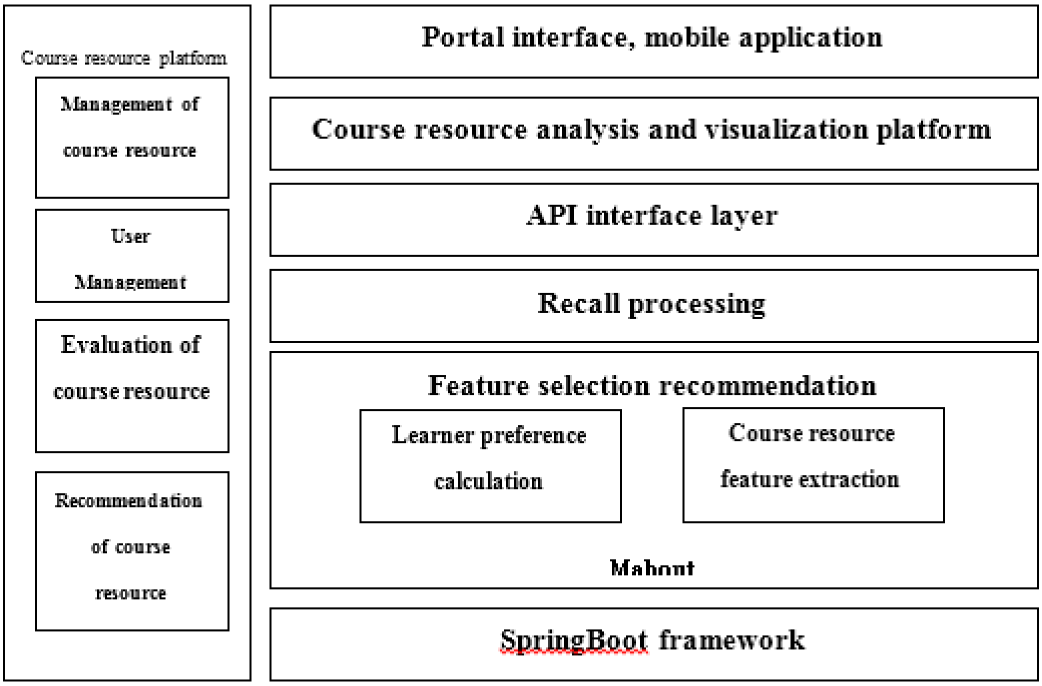

Figure 2 Implementation process of course resource recommendation system

First and foremost, define the data model of Mahout, which provides a variety of DataModels, such as MysqlJDBCDataModel, for the operation of the mysql database, HBaseDataModel, for the operation of the hbase database, FileDataModel, for the operation of file storage data (such as csv), etc. .

Then, based on the learner behavior analysis, a learner preference model is established. The learner preference label calculation formula: learner label weight = behavior type weight × time decay × learner behavior times × TF-IDF calculation on label weight. Among them, behavior type weights: different behaviors such as browsing, searching, collecting, ordering, and purchasing of learners have different significance to learners. Generally speaking, the higher the operation complexity, the greater the behavior weight. The weight value is generally given subjectively by the data analyst; time decay: some behaviors of the learner are weakened by the influence of time, the farther the behavior time is from now, the less meaning the behavior has for the learner at present; behaviors time: the weight of a learner's label is counted by day. The more times a learner acts with the label on a certain day, the greater the impact of the label on the learner; TF-IDF calculation on label weight: the product of multiplying the importance of each label to the learner and the importance of the labels in overall is used to obtain the objective weight value of each label. For the sake of calculating the learner preference label, the corresponding weight value of the learner behavior label needs to be calculated on the basis of the learner behavior label. The aggregated weights to the similar labels calculate learner preference label.

Finally, Generic User-Based Recommender is defined with reference to different methods, for instance, Generic User-Based Recommender is based on the learner recommendation. By calling the recommended (long userID, int howMany) method, insert the id of the learner and the number of return recommendation.

(6) Conclusion

This research focuses on personalized recommendation algorithms and proposes a feature selection-based recommendation algorithm. A feature matrix is generated by extracting the features of items. At the same time, a learner preference matrix is generated based on the calculation of learner's interest biases over a period of time. As a result, a feature selection-based recommendation algorithm is implemented. The feature recommendation course resource platform is utilized. The model establishment and system implementation process is discussed in detail. It is hoped to provide reference ideas for the construction of university course..

Acknowledgement

Construction and research project of teaching resource cloud platform under the background of "big data" of "13th five year plan" educational research project of Guangdong Education Society.(Project number:GDES13614); the first batch of reform of innovation and business running education with cooperation between production and learning of the Ministry of Education in 2018 (project number: 201801193131); and the first batch of reform of teaching content and curriculum system of cooperation between production and learning of the Ministry of Education in 2018 (project number: 201801193063); the cooperative education project based on school-enterprise cooperation in Guangdong Province in 2018 (project number: PROJ999608519785320448);

References

[1] Monica Rogati, Yiming Yang: High-performing feature selection for text classification[J], ProceedingCIKM '02 Proceedings of the eleventh international conference on Information and knowledge management ,Pages 659-661

[2] Good N, Schafer J B, Konstan J A, et al. Combining collaborative filtering w ith personal agents for better recommendations[C]/ /Proceedings of the Sixteenth National Conference on Artificial In-telligence(AAAI) , 1999: 439-446.

[3] YangWenfen, LiXing. Chinesekeywordextractionbasedonmax-duplicatedstringsofthedocuments [C] //Proceedingsofthe25th AnnualInternationalACM SIGIRConferenceonRe-SearchandDevelopmentinInformationRetrieval, 2002

[4] AGICHTEINE, CASTILLO C, DONATO D, et al. Finding high-quality content in social media [C] / /Association for Computing Machinery. Proceedings of the 2008 international conference on web search and data mining. ACM, 2008:183-194.

[5] Zhang Jun, Liu Man, Peng Wei-ping, et al. Collaborative filtering recommendation algorithm based on fusion interest and score[J].Journal of Chinese Computer Systems, 2017, 38(2) : 357-362.Zhang Jun, Liu Man, Peng Wei-ping, et al. Collaborative filteringrecommendation algorithm based on fusion interest and score[J].Journal of Chinese Computer Systems, 2017, 38(2) : 357-362.

[6] Jiang M, Cui P, Liu R, etal. Social contextual reco-mmendation[C]/ /Proceedings of the 21st ACM International Conference on Information and Know ledge M anagement (CIKM), 2012: 45-55.

[7] Bridge D., Kelleher J. Experiments in sparsity reduction: usingclustering in collaborative recommenders [C]. Proceedings of the 13th IrishInternational Conference on Artificial Intelligence and Cognitive Science.London:Springer-Verlag,2002:144-149.

[8] Kelleher J., Bridge D. Rectree centroid: an accurate,scalablecollaborative recommender [C]. Proceedings of the 14th Irish Conference onArtificial Intelligence and Cognitive Science, 2003:89-94.

[9] Huang Z., Chen H., Zeng D. Applying associative retrieval techniques toalleviate the sparsity problem in collaborative filtering [C]. ACM Transaction onInformation Systems, 2004, 22(1): 116-142.

[10] HuY H,Chen Y L.Mining association rules with multiple minimum sports: a new mining algorithm and a support tuning mechanism[J].2004 ,42(1):1-24.

[11] DongLiang XU.Research on clustering algorithm based on Text Mining [J].Microcomputer information, 2011, 27(2): 168-169.

[12] Manying Zhou.On Chinese word segmentation technology of Baidu and Google [J].China

Index, 2011, 9(2): 44-46.

[13] Mingbian Ren.The method and practice of intelligent information system optimizing data modeling by association knowledge: Zhejiang University Press, 2012.03

[14] https://blog.csdn.net/Wetsion/article/details/80160706

[15] Hang Yin. Research on collaborative filtering technology in information recommendation system [D]. Northeastern University,2012.

[16] Chenyuan Yang. Personalized advertising recommendation based on commodity category [D]. Xi'an University of Electronic Science and technology,2017.