**Received:** January 10, 2018
**Revision received:** June 2, 2018
**Accepted:** June 5, 2018


**www.estp.com.tr**
**DOI** 10.12738/estp.2018.5.094 ♦ October 2018 ♦ 18(5) ♦ 1948-1959

***Research Article***

# Research on Personalized Recommendation of Educational Resources Based on Big Data*

Dewen Seng[1,2]
*Hangzhou Dianzi University; Key Laboratory of Complex Systems Modeling and Simulation*

Xiuli Chen[1,2]
*Hangzhou Dianzi University; Key Laboratory of Complex Systems Modeling and Simulation*

Xujian Fang[1,2]
*Hangzhou Dianzi University; Key Laboratory of Complex Systems Modeling and Simulation*

Xuefeng Zhang[1,2]
*Hangzhou Dianzi University; Key Laboratory of Complex Systems Modeling and Simulation*

Jing Chen[1,2]
*Hangzhou Dianzi University; Key Laboratory of Complex Systems Modeling and Simulation*

## Abstract

With the rapid development of Internet technology, the era of education informationization has arrived. With massive education resources, users are faced with the problem of information overload. This essay takes the management of educational resources and the big data which form the platform as the background, and designs a personalized recommendation algorithm of educational resources according to the users' own characteristics. Personalized recommendation algorithm of EDX platform and personalized paper recommendation of CiteULike are used to verify the rationality and effectiveness of the proposed algorithm.

## Keywords

Big Data • Educational Resources • Personalized Recommendation

*This work is supported by the National Natural Science Foundation of China (No. 61473108), the Public Projects of Zhejiang Province (No. 2013C33082), Zhejiang Provincial Natural Science Foundation (No. LY15F020038 and No. LQ13F020005) and the research foundation of the Education Department of Zhejiang Province (No. Y201430884).
[1]**Correspondence to:** Dewen Seng, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 10018, China. Email: sengdw@163.com
[2]Key Laboratory of Complex Systems Modelling and Simulation, Ministry of Education, Hangzhou 10018, China.

**Citation:** Seng, D. W., Chen, X. L., Fang, X. J., Zhang, X. F., Chen, J. (2018). Research on Personalized Recommendation of Educational Resources Based on Big Data. *Educational Sciences: Theory & Practice, 18(5),* 1948-1959. http://dx.doi.org/10.12738/estp.2018.5.094

In the era of Internet education era, large-scale Internet education resources have come into being, in which the most representative one is the massive open online courses (MOOC), attracting millions of registered users, and their learning behavior can be regarded as the learning big data.

In the face of massive educational resources, the majority of users choose learning content through keyword searching. However, due to the information overload, this measure is restricted by many factors. Users may not find their desired learning resources.

This essay takes the management of educational resources and the big data which form the platform as the background (GarcíA & GarcíA, 2005; Ponsard, Touzani, & Majchrowski, 2018). Through analyzing learning the characteristics of the user, the user item rating matrix, and based on LDA topic model, collaborative recommendation algorithm (Calverley & Shephard, 2003; Constantin, Dahimene, Mouza, & Grossetti, 2016), the personalized recommendation algorithm of educational resources has been designed. It can help users get more learning resource recommendations after entering what they want to search the theme, so that the users can find relevant resources in a more convenient way. The essay takes the personalized education resources recommendation on edx platform and the individualized paper recommendation on CiteULike platform as examples. Through comparing these to the other two kinds of recommendation algorithms, this paper verified the rationality and validity of Topic-Based CF.

## Design of personalized recommendation algorithm of educational resources

### Related technologies

**User-item rating matrix.** Many recommendation algorithms are based on the user item rating matrix, and so do the algorithms in this essay (Grigal & Others, 1997). Its general form is as follows:

$$\begin{pmatrix} a11 & \cdots & a1n \\ \vdots & \ddots & \vdots \\ am1 & \cdots & amn \end{pmatrix} \tag{1}$$

among them: $a_{ij}$ indicates the rating of item i for user i

The matrix can usually be rated by explicit or implicit ways (Shriner & Destefano, 2003). The implicit rating method, because of its wide application range, is commonly used. Implicit rating method can be divided into mean value method and collaborative filtering based rating method. Based on the mean value method, this paper puts forward the user-item implicit evaluation method.

**LDA topic model.** LDA is a document theme generation model, which is composed of three layers (Prettifrontczak & Bricker, 2000). Word is the smallest structural unit; theme is composed of words; document can be regarded as the set of articles. The so-called generative model (Espin, Deno, & Albayrakkaymak, 1998) refers to the probability distribution of the upper structure of the word, theme and document. In this paper, we use the model to classify the themes in the educational resource database, and figure 1 is the graph model of the LDA model (Smith, 1990).

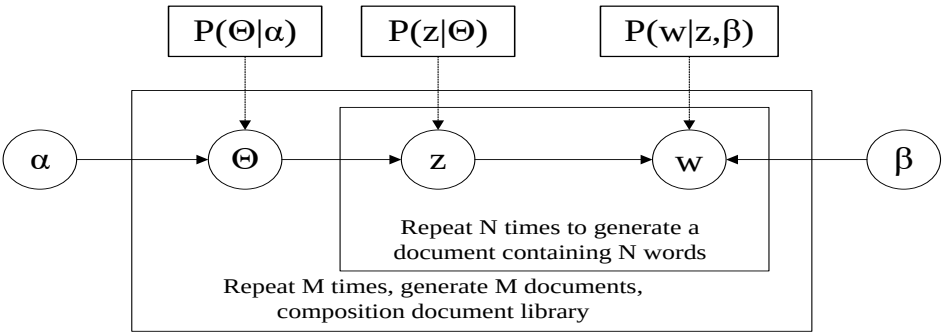


*Figure 1*. LDA theme model

Note: α: Distribution vector parameter; Θ: theme vector; z: subject variable; β: the probability density of the word corresponding to the subject z under the condition; w: word variable.

## Collaborative filtering recommendation algorithm

Collaborative filtering recommendation algorithm (Yates *et al*., 2011) identifies and recommends relative resources may interest the users based on users behavior and their interests. Among those earliest recommendations used in e-commerce, Amazon is a typical application of collaborative filtering algorithm which is now widely used in various fields. The user-item rating matrix mentioned above is the basis of this recommendation algorithm.

Collaborative filtering recommendation algorithms can be divided into three categories (Coifman & Wickerhauser, 1992). Since the number of users of educational resources is larger than the amount of resources, this paper is based on the item collaborative filtering recommendation algorithm, and the detailed process is shown in Fig. 2.

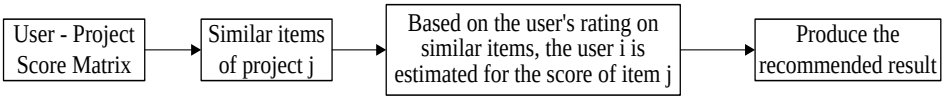


*Figure 2*. Item - based collaborative filtering recommendation algorithm

## User-item implicit estimation

The educational resources personalized recommendation algorithm in this paper is based on user-item rating matrix. Based on previous rating method, the paper designs an implicit estimation method, which can reflect the users' learning characteristics, and introduces the concept of information entropy and the forgetting curve to distribute of user behavior and time weight (Dolog, Simon & Nejdl, 2008).

This paper analyzes the courses and relative data provided by Berkeley University and Harvard University on edx platform in 2015-2016. Fig. 3 is an example of the EDX platform. According to the behavior of four

different types of users (registered users, users who have completed the courses, users who are browsing, and users who are learning the user with different behavior) (Köck & Paramythis, 2011), user-item implicit rating estimation method is established (Yu, 2012). As shown in Figure 4, user behaviors include the frequency and the duration of the interactions between the user and the curriculum. The specific process is shown in Fig. 5.

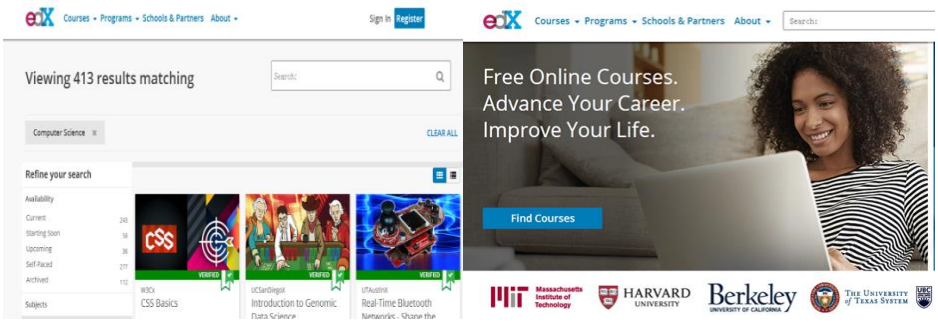


*Figure 3.* EDX course platform example

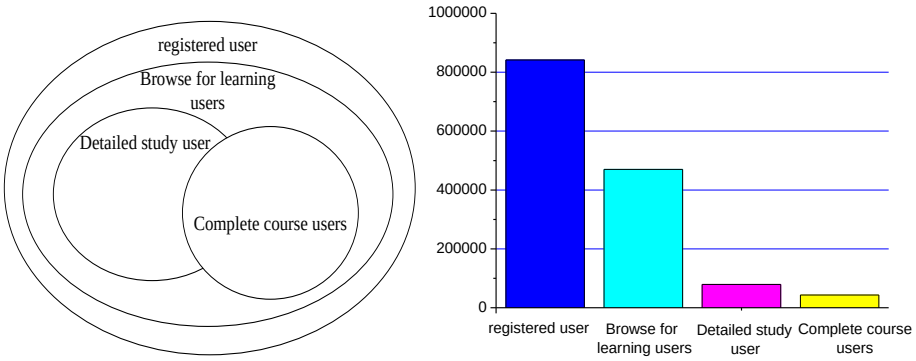


*Figure 4.* Four users' inclusion relationships and statistics

Fig. 4 shows the containment relationships between different users, which shows that although there are a large number of users (about 850000) have registered the online learning platform, half of them are browsing (only about 45000 people have finished learning, the number of users who are learning is less than 80000).

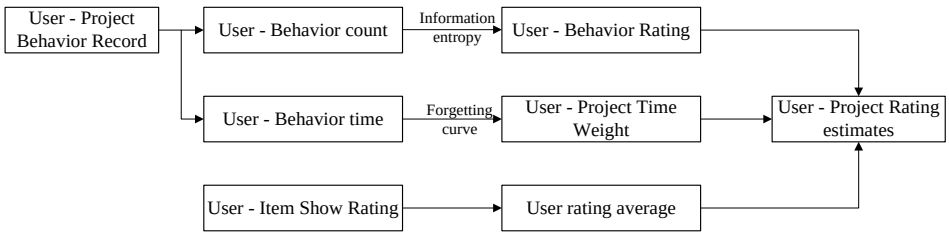


*Figure 5.* User - Item Rating Estimation Process

**User interest model based on topic model**

Analyze hash-tag (system information and user-defined label) under educational resources based on LDA topic model; establish topic model by combining project - evaluation matrix; estimate parameters under topic model by adopting parameter estimation methods of Gibbs sampling (Chen, Lee & Chen, 2005); then build up user interest model based on topic model with the guidance of user project rating matrix (Xu, Wang & Wang 2005). Please check Image 6 for main process. This model will show users' preferences on learning better, and will reduce the similarity calculation under recommendation algorithm.

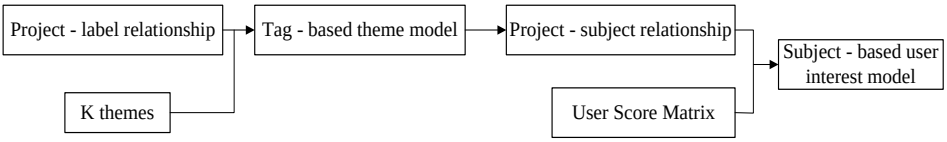


*Figure 6.* Subject - based user interest model

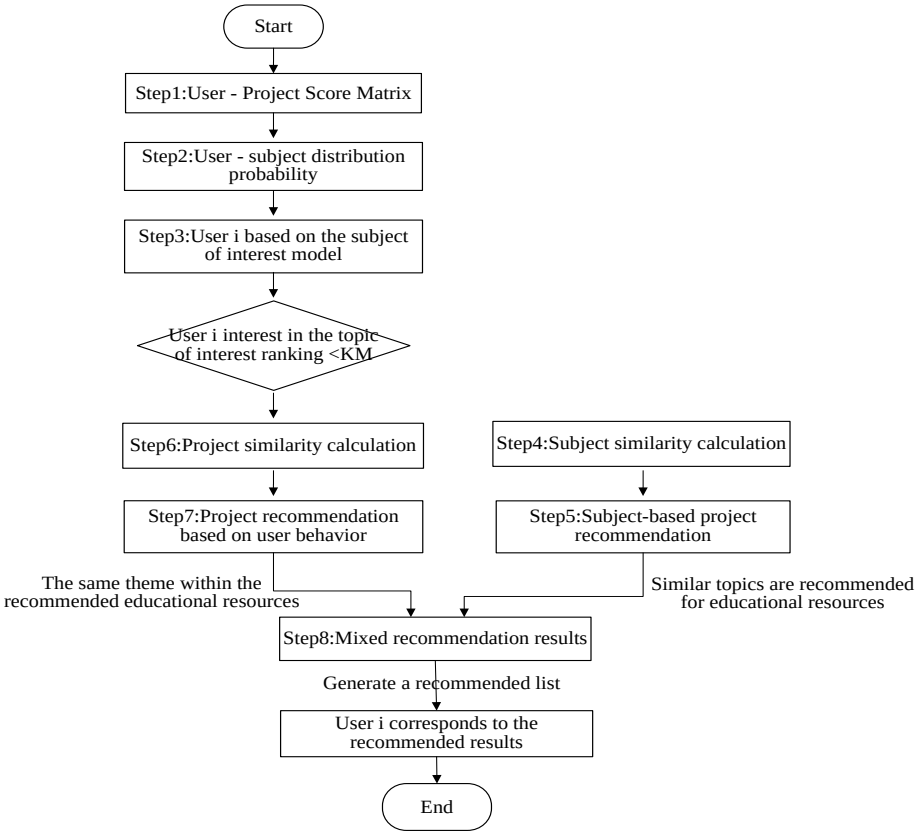


*Figure 7.* An algorithmic flow chart for recommending educational resources for user i

**Process of personalized recommendation algorithm (topic-based CF) for educational resources**

Based the above-mentioned LDA topic model, recommendation algorithm of collaborative filtering, user - project implicit rating estimation, and topic-based users' interest model, personalized recommendation algorithm (Topic-Based CF) for educational resources is presented (Hsu, 2008); please check Image 7 for concrete process. The algorithm takes collaborative filtering as basic idea, recommends results of behavior-based items recommendation and results of topic-based items recommendation to gain recommendation list, then get recommendation result. Detailed process of recommendation is generated by doing cloud computing based on big data formed by educational resources platforms. It will not be discussed in detail here.

## Case analysis on personalized recommendation of educational resources

**Evaluation methods**

The paper will adopt three indicators, i.e. accuracy rate (Formula 2), recall rate (Formula 3) and F1 score (Formula 3), to evaluate recommendation results.

$$Precision = \frac{\sum u \in U|R(u) \cap I(u)}{\sum u \in U|R(u)|} \quad (2)$$

$$Recall = \frac{\sum u \in U|R(u) \cap I(u)}{\sum u \in U|I(u)|} \quad (3)$$

$$F1\text{score} = \frac{2 * Pr\,ecision * Recall}{Pr\,ecision + Recall} \quad (4)$$

among them: u: user; U: User collection; R(u): The recommended item collection for the user u recommendation algorithm; I(u): A collection of items that the user actually generates behavior

Adopt Topic-Based CF recommendation algorithm and other two recommendation algorithms (Item-Based CF algorithm and Behavior-Based CF algorithm) to select N, the length of different recommendation lists; compare the accuracy rate of N and F1 score value, then verify the effectiveness and rationality of Topic-Based CF algorithm presented by the paper.

Table 1
*EDX Data Description*

| Serial number | Type of data | Content |
|---|---|---|
| 1 | String | Course ID |
| 2 | String | User ID |
| 3 | Boolean | Whether the user is registered |
| 4 | Boolean | Whether the user has studied more than 1/2 of the course section |
| 5 | Boolean | Whether the user has completed the study |
| 6 | Float | The user-course score is represented by 0-1 |
| 7 | Time Stamp | Start learning time |
| 8 | Time Stamp | Last login time |
| 9 | Int | The number of days of interaction with the course |
| 10 | Int | Watch the number of video resources |
| 11 | Int | Participate in the discussion sessions in the course forum |

**Personalized course recommendation of educational resources based on EDX platform**

**Experimental Data.** Select data about users' behaviors and courses opened by Harvard University and Berkeley College on edx platform during the academic year of 2015 - 2016; it totally contains 68, 577KB, 641321 data, and each data includes contents shown on Table 1.

**Analysis on experimental results.** According to methods mentioned above, the paper firstly does weight analysis on data of users' behaviors by adopting weight allocation methods of information entropy; results are shown on Image 8. Then the paper obtains user - course score matrix based on the results of weight distribution.

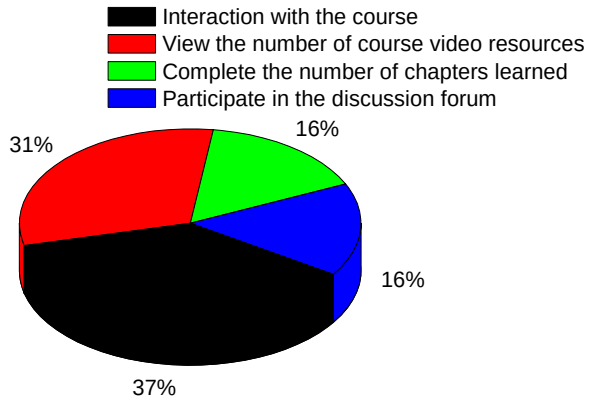


*Figure 8.* User Behavior Weight

In edx data set, topics of courses do not have explicit tags, but they could be gained by checking ID of course; for example, topic classification of Harvard University on edx website is Computer Science; course under Computer Science is Css Basics, and its corresponding ID is HarvardX/CB21X/2015_autumn. Therefore, it could be deemed that distribution probability between topic and course is 1.

Distribute data set at random; among which, training set occupies 2/3, and test set holds 1/3. Based on user - course score and course - topic distribution, the paper adopts process of personalized recommendation algorithm for educational resources which is specially designed under the paper to get the recommendation list shown on Table 2; on the left is users' ID information, and on the right is the information of recommended course ID.

Table 2
*Examples of recommended results*

| User ID address | Recommended course ID address |
|---|---|
| 130364489 | 6.00x |
| 130447948 | 5.02x,6.00x |
| 130491149 | 6.00x |
| 130457959 | 5.02x,8.HRev,7.00x,3.092x, 6.00x,14.72x |
| 130265316 | 6.00x |
| 13024568 | 8.02x,6.00x |
| 13055502 | 8.02x,6.00x |
| 130079457 | 8.02x8.HRev, 7.00x3.091x,6.00x,14.72x |
| 13011546 | 6.00x |
| 13004867 | 6.02x |

Adopt selection method of TOP-N, which is select top N from all recommended results as the representative; as N changes, accuracy rate, recall rate and F1 score of algorithm change accordingly. Image 9 and Image 10 are the comparison of accuracy rate and F1 score of three algorithms when N changes.

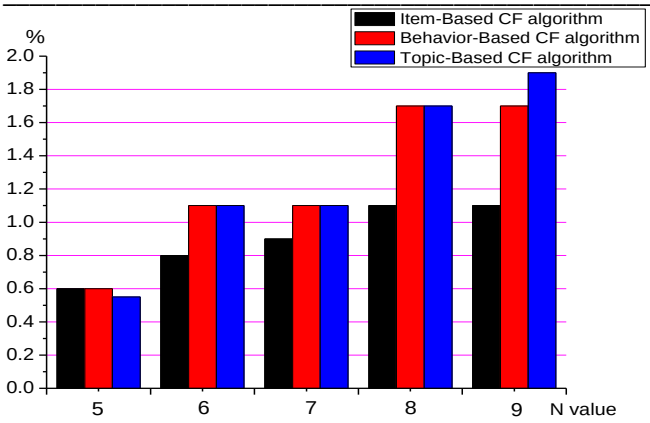


*Figure 9.* Comparison of accuracy

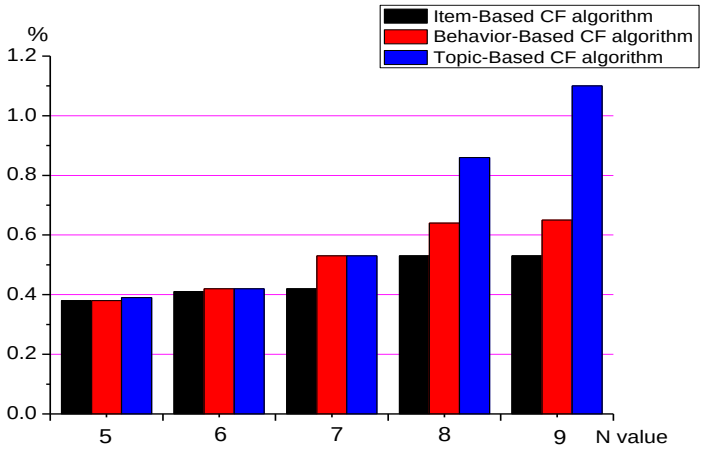


*Figure 10.* F1 score comparison

Analysis on recommended results of personalized course of educational resources on edx platform:

(1) Compared with other two algorithms, accuracy rate of Topic-Based CF recommendation algorithm presented by the paper is improved when N is added; and the increasing amplitude is obviously larger than the ones of the other two algorithms.

(2) F1 score value of Topic-Based CF recommendation algorithm is obviously better than the one of the other two algorithms. F1 score value could be used to comprehensively evaluate the accuracy rate and recall rate of algorithm, therefore, it could be found that effectiveness of Topic-Based CF recommendation algorithm presented by the paper is superior to the ones of the other two algorithms.

(3) Based on two indicators, i.e. accuracy rate and F1 score value, accuracy and effectiveness of recommended results of personalized course of educational resources on edx platform could be obtained. It also indicates that user - project score estimation and topic-based users' interest model presented by the paper meet with actual demand.

**Personalized paper recommendation on CiteuLike platform**

**Experimental data.** CiteULike is a platform which has various kinds of paper resources and combines paper presentation, retrieval, and user-defined tag together. Please check Image 11; users are able to use such platform to store, manage and share their papers.

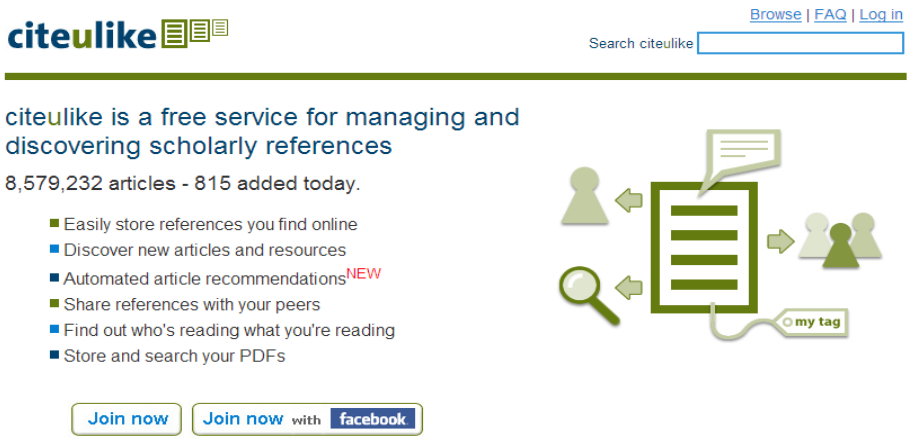


*Figure 11.* CiteULike Paper resource platform

The paper select data of 5673 users, 16988 paper resources, 203988 users' behavior records about paper projects, and 16983 user-defined tags from CiteULike as experimental data. Please check Table 3 for specific contents of data.

Table 3
*A Description of the Experimental Data Set*

| File name | Data description |
|---|---|
| Tag.ietm.dat | Tag - the paper data, each row of data including the paper ID and the corresponding paper label information |
| Rawtext.dat | The original data set includes the full text of the captured paper, the label, and the user data |
| Tags.dat | The tag data set, including the tag ID and the corresponding tag word |
| Users.dat | User - The paper shows the score matrix |

**Analysis on experimental results.** The paper divides user - paper behavior data into test set and training set at random; then the paper adopts process of personalized recommendation algorithm (Topic-Basic CF) for educational resources presented under this paper to firstly estimate user - paper score, then establish user - topic model and user - paper score matrix, and finally gain recommended results based on related algorithm.

Similar to experiment about recommendation of personalized course of educational resources on edx platform, the paper will compare accuracy rate (Image 12) and F1 score (Image 13) of three algorithms when N changes.

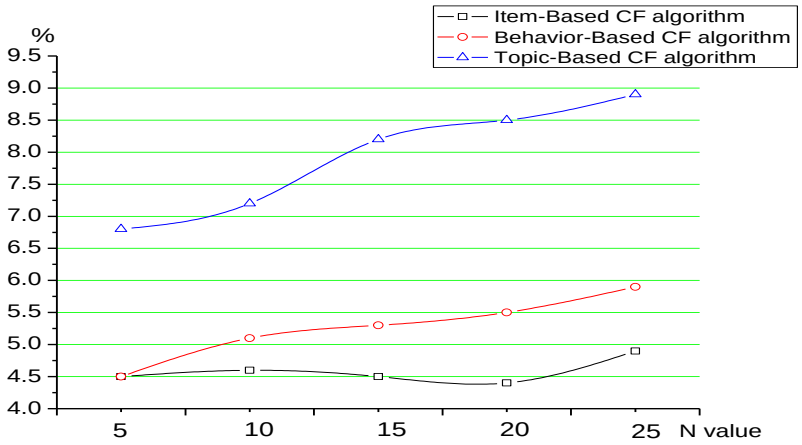


*Figure 12.* Comparison of accuracy

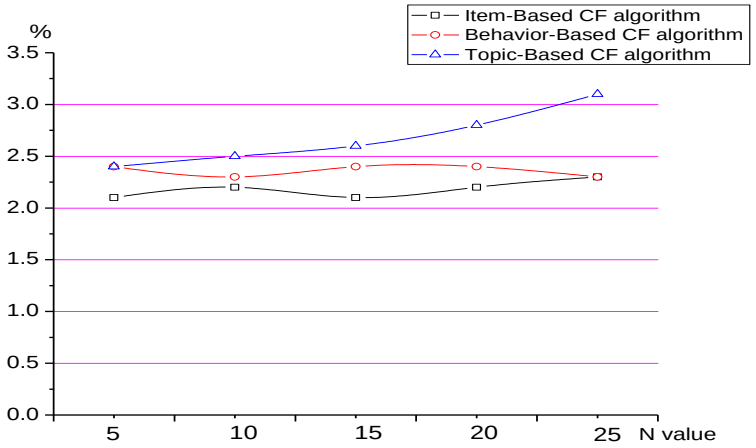


*Figure 13.* F1 score comparison

Analysis on recommended results of personalized papers on CiteUlike platform:

(1) Among three algorithms, value of accuracy rate and F1 score of recommendation algorithm (Topic-Based CF) presented by the paper increase when value of N increases; and the accuracy rate of Topic-Based CF is obviously higher than the ones of the other two algorithms. It indicates that the introduction of topic-based project recommendation model will greatly improve the accuracy rate of recommended results.

(2) Comparing Image 13 with Image 14, it could be found that accuracy rate of Topic-Based CF recommendation algorithm is far higher than F1 score; according to analysis on the above-mentioned Formula (4), F1 score is affected by both accuracy rate and recall rate; therefore, it could be found that recall rate of Topic-Based CF recommendation algorithm does not increase obviously as value of N increase; so further optimization could be done to improve F1 score of Topic-Based CF recommendation algorithm.

(3) According to experimental results, it could be found that effectiveness is obvious when topic-based user interest model is introduced into Topic-Based CF recommendation algorithm.

## Conclusion

The paper takes big data generated by online educational resources platforms such as MOOC as the background, analyzes users' characteristics of learning and behaviors, and combines related skills of previous recommendation algorithms to design personalized recommendation algorithm for educational resources (Topic-Based CF); then the paper takes personalized course recommendation for educational resources on edx platform and personalized paper recommendation on CiteULike platform as the examples to verify the effectiveness and rationality of Topic-Based CF, and to gain the following achievements:

(1) Taking previous recommendation algorithms for reference, the paper designs project - implicit rating estimation, and introduces concepts of forgetting curve and information entropy to distribute user behavior weight reasonably.

(2) Design user interest model of topic model based on LDA topic model; this model obviously improves the accuracy rate and F1 score of Topic-Based CF recommendation algorithm.

(3) Do experiment about personalized course recommendation for educational resources on edx platform and experiment about personalized paper recommendation on CiteULike to verify the advantage and rationality of Topic-Based CF recommendation algorithm.

(4) Recall rate of personalized recommendation algorithm (Topic-Based CF) for educational resources is relatively low; it shall start with recall rate to further improve and perfect Topic-Based CF recommendation algorithm.

(5) Based on personalized recommendation algorithm (Topic-Based CF) for educational resources, users will be able to find needed resources easily, quickly and accurately; purpose of teaching students in accordance with their natural abilities will be finally realized.

## References

Calverley, G., & Shephard, K. (2003). Assisting the uptake of on-line resources: why good learning resources are not enough. *Computers & Education, 41(3)*, 205-224. https://dx.doi.org/10.1016/S0360-1315(03)00028-9

Chen, C. M. (2009). Personalized e-learning system with self-regulated learning assisted mechanisms for promoting learning performance. *Expert Systems with Applications, 36(5)*, 8816-8829. https://doi.org/10.1109/ICALT.2007.205

Chen, C. M., Lee, H. M., & Chen, Y. H. (2005). Personalized e-learning system using item response theory. *Computers & Education, 44(3)*, 237-255. https://doi.org/10.1016/j.compedu.2004.01.006

Coifman, R. R., & Wickerhauser, M. V. (1992). Entropy-based algorithms for best basis selection. *IEEE*

*Transactions on Information Theory, 38(2)*, 713-718. https://dx.doi.org/10.1109/18.119732

Constantin, C., Dahimene, R., du Mouza, C., Grossetti, Q. (2016). User-based recommendations for micro-blogging systems, *Ingenierie des Systemes d'Information, 21(3)*, 93-118. https://doi.org/10.3166/ISI.21.3.93-118

Dolog, P., Simon, B., & Nejdl, W. (2008). Personalizing access to learning networks. *ACM Transactions on Internet Technology*, 8(2), 3. https://dx.doi.org/10.1145/1323651.1323654

Espin, C. A., Deno, S. L., & Albayrakkaymak, D. (1998). Individualized education programs in resource and inclusive settings: how "individualized" are they?. *Journal of Special Education, 32(3)*, 164-174. https://dx.doi.org/10.117/002246990803200303

GarcíA, F. J., & GarcíA, J. (2005). Educational hypermedia resources facilitator. *Computers & Education, 44(3)*, 301-325. https://dx.doi.org/10.1016/j.compedu.2004.02.004

Grigal, M., & Others, A. (1997). An evaluation of transition components of individualized education programs. *Exceptional Children, 63(3)*, 357-372. https://dx.doi.org/10.1177/001440299706300305

Hsu, M. H. (2008). A personalized English learning recommender system for esl students. *Expert Systems with Applications, 34(1)*, 683-688. https://dx.doi.org/10.1016/j.eswa.2006.10.004

Köck, M., & Paramythis, A. (2011). Activity sequence modelling and dynamic clustering for personalized e-learning. *User Modeling and User-Adapted Interaction, 21(1-2)*, 51-97. https://dx.doi.org/10.1007/s11257-010-9087-z

Ponsard, C., Touzani, M., & Majchrowski, A. (2018). How to conduct big data projects: Methods overview and industrial feedback, *Ingenierie des Systemes d'Information, 23(1)*, 9-33. https://dx.doi.org/10.3166/ISI.23.1.9-33.

Prettifrontczak, K., & Bricker, D. (2000). Enhancing the quality of individualized education plan (iep) goals and objectives. *Journal of Early Intervention, 23(2)*, 92-105. https://dx.doi.org/10.1177/105381510002300204

Shriner, J. G., & Destefano, L. (2003). Participation and accommodation in state assessment: the role of individualized education programs. *Exceptional Children, 69(2)*, 147-161. https://dx.doi.org/10.1177/001440290306900202

Smith, S. W. (1990). Comparison of individualized education programs (IEPS) of students with behavioral disorders and learning disabilities. *Journal of Special Education, 24(1)*, 85-100. https://dx.doi.org/10.1177/002246699002400107

Xu, D., Wang, H., & Wang, M. (2005). A conceptual model of personalized virtual learning environments. *Expert Systems with Applications, 29(3)*, 525-534. https://dx.doi.org/10.1016/j.eswa.2005.04.028

Yates, T., Davies, M. J., Sehmi, S., Gorely, T., & Khunti, K. (2011). The pre-diabetes risk education and physical activity recommendation and encouragement (prepare) programme study: Are improvements in glucose regulation sustained at 2years?. *Diabetic Medicine, 28(10)*, 1268-1271. https://dx.doi.org/10.1016/j.pec.2008.06.010

Yu, S. J. (2012). The dynamic competitive recommendation algorithm in social network services. *Information Sciences, 187(1)*, 1-14. https://dx.doi.org/10.1016/j.ins.2011.10.020