Received: January 7, 2018
Revision received: May 18, 2018
Accepted: May 25, 2018

www.estp.com.tr
DOI 10.12738/estp.2018.6.191 ⬩ December 2018 ⬩ 18(6) ⬩ 2922-2931

Short Communication

Topic Interaction Model Based on Local Community Detection in MOOC Discussion Forums and its Teaching Application *

Gengxin Sun[1]
Qingdao University

Sheng Bin[2]
Qingdao University

Abstract

Discussion forums is a major component of MOOC platform. The teaching interaction in MOOC learning process mainly occurs in discussion forums by topic. Aiming at the problem of students' ability difference in the course of MOOC teaching, in this paper a topic interaction model based on local community detection is proposed. Through topic modeling for students' interaction ability in discussion forums, local community detection algorithm is used to classify students' various abilities reasonably. Through tracking and analyzing student behavior information on a highly interactive MOOC platform, the accuracy rate of the proposed model is obviously higher than that of the traditional assessment methods. The teaching practice using the model shows that students' abilities in all aspects are improved by means of pertinence classroom communication and training.

Keywords

Topic Interaction • MOOCs • Topic Model • Community Detection

*This work is supported by the Humanity and Social Science Youth foundation of Ministry of Education of China (grant no.15YJC860001). This research is also supported by the Natural Science Foundation of Shandong Province (grant no. ZR2017MG011), Shandong Education Science 13th Five-Year plan project (grant no. BYK2017006), Statistical Science Research Project of China (grant no. 2015LZ20, 2017LZ38) and Humanities and Social Sciences Foundation of Shandong Province (grant no.17CHLJ16), Humanities and Social Sciences Foundation of Qingdao (grant no. QDSKL1701074).
[1]**Correspondence to:** Gengxin Sun (PhD), School of Data Science and Software Engineering, Qingdao University, Qingdao 266071, China. Email: sungengxin@qdu.edu.cn
[2]School of Data Science and Software Engineering, Qingdao University, Qingdao 266071, China. Email: binsheng@qdu.edu.cn

Citation: Sun, G. X., & Bin, S. (2018). Topic Interaction Model Based on Local Community Detection in MOOC Discussion Forums and its Teaching Application. *Educational Sciences: Theory & Practice, 18(6),* 2922-2931. http://dx.doi.org/10.12738/estp.2018.5.191

MOOC is an online course model that has sprung up in recent years (Ma, Zheng, & Zhao, 2015; Chen, Davis, Lin, Hauff, & Houben, 2016). Unlike the traditional online teaching, MOOC combines online comprehensive learning management with more open network resources in a brand-new form of knowledge diffusion, and it pays more attention to the interaction between teachers and students. At present, MOOC is attracting wide attention from educational institutions, especially the higher education colleges in various countries. The problems of various online support technologies and management modes have become a research hot issues.

Unlike the traditional distance education, open online courses and online applications based on the network, the interaction of MOOC is stronger. Under the model of MOOC, course selection, classroom teaching, students' learning process, teacher-student interaction process and so on are fully and systematically realized in MOOC platform. Most of the online interaction between teachers and students is carried out through the discussion forum (Konstan, Walker, Brooks, Brown, & Ekstrand, 2015), which plays an indispensable role in the whole teaching process and knowledge transfer process. However, since the course of MOOC relies heavily on self-study, it requires learners to have strong self-study ability and interaction ability, but the level of students is uneven, it will greatly affect the teaching effect of MOOC. Therefore, how to classify the different levels, and according to students' needs and level how to determine teaching objectives has become the primary problem which restricted the development of MOOC. This is a question of teaching reform, and also a technical problem.

In this paper, a topic interaction model based on local community detection (LCD) is proposed for MOOC. Taking "computer programming language" course as an example, through accurate platform user tracking, topic model information collection and user behavior evaluation based on LCD algorithms, the learning situation and the interaction in the discussion forum of each user were fully understood and mastered. It will provide a good online platform and technical support for improving the quality of student training.

Preliminaries

Topic model

Garrison, Anderson, & Archer (1999) proposed a teaching transaction model to provide reference for exploring the construction of MOOC discussion forum. The model is shown in Figure 1.

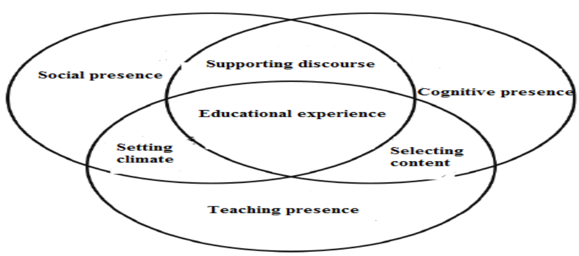

Figure 1. Elements of educational experience in the model

The model defines cognitive presence as the degree to which learners can construct meaning through continuous communication in any discussion forum. Teaching presence refers to the provision of subject knowledge in professional fields and the promotion of active learning through the design and management of learning sequences. In order to study the relationship between topic model and discussion forums, Blei, Ng, & Jordan, (2012) proposed latent dirichlet allocation model, it is a three-level Bayesian probability model, which is divided into three layers: word, topic and document. The model is shown in Figure 2.

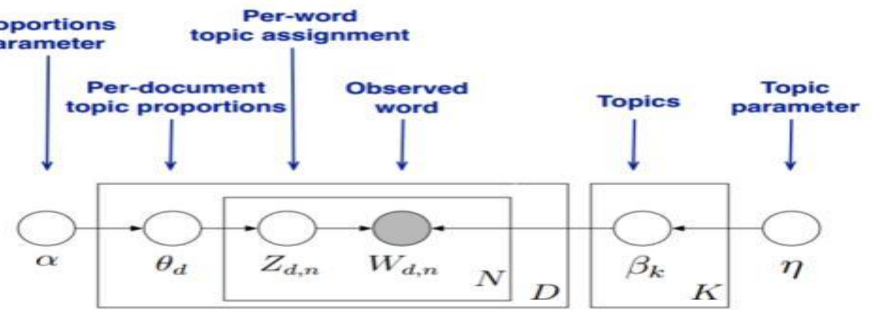

Figure 2. Structure diagram of LDA topic model

The model can be used to detect potential topic information in large scale document sets or forums. As the most effective topic model at present, LDA has been successfully applied to a large number of text related fields (Jian-Hua, & Zheng-You, 2014; Li, Zhou, Sun, & Zhang, 2016; Mothe, & Rakotonirina, 2018; Liang, Liu, Tan, & Bai, 2014).

In the LDA model, if the number of topics is *K* and the number of words is *V*, then each document is considered to be a mixture of *K* different topics. Its distribution can be expressed as:

$$\theta = [P(z_1), P(z_2), \dots, P(z_2)]$$

$$\theta \sim p(\theta) \sim \mathrm{Dir}(\alpha)$$

The probability distribution of *V* words in each topic can be expressed as $P(\mathrm{w}|\mathrm{z})$, it can be defined by $K \times V$ matrix as follows:

$$\begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_K \end{bmatrix} = \begin{bmatrix} P(w_1|z_1) & \cdots & P(w_V|z_1) \\ \vdots & \ddots & \vdots \\ P(w_1|z_K) & \cdots & P(w_V|z_K) \end{bmatrix}$$

$$\varphi \sim p(\varphi) \sim \mathrm{Dir}(\beta)$$

Each row in the matrix represents the probability distribution of all words in topic z_i. At present, Gibbs sampling structure (Gelfand, 2000) is used to converge to a Markov chain of target probability distribution. The samples closest to the probability distribution value are extracted to indirectly estimate θ and φ.

For generating the LDA model, a topic number $w_i \in z_i, i = 1,2, \dots, V$is randomly assigned to each word of the *m*-th document in the corpus. Then the topic of w_i is re-sampled according to Gibbs sampling method.

$$p(z_k|Z_{\neg i}, w) \propto \hat{\theta}_{mk} \cdot \hat{\varphi}_{kt} = \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_{m,\neg i}^{(v)} + \alpha_k)} \cdot \frac{n_{k,\neg i}^{(v)} + \beta_v}{\sum_{v=1}^{V}(n_{k,\neg i}^{(v)} + \beta_v)}$$

where $\neg$ represents w_i is not included. $n_{m,\neg i}^{(k)}$ represents the number of topics in the *m*-th document. $n_{k,\neg i}^{(v)}$ represents the number of words in *k*-th topic. The above sampling process is repeated until converges of Gibbs sampling.

Local community detection

Nodes in complex networks can be grouped with dense intra-group links and sparse inter-group links (Palla, Derényi, Farkas, & Vicsek, 2005). The sub-network consisting of nodes in the same group and links between nodes is called community in complex networks. Because of the large scale of the network in real life, it is difficult to obtain the community structure of all nodes, and people tend to pay more attention to the local community structure of some nodes. So we need to use a community detection algorithm which only considers the local network topology structure (Seth, Bhattacharyya, & Kim, 2014; Moctar, & Sarr, 2016), one or more seed nodes are specified firstly, and then the natural communities which these nodes belong to are mined.

Traditional local community detection algorithms include neighborhood-based merging framework algorithm and random walk-based algorithm (Thakur, Tiwari, Thai, Chen, & Dress, 2009). Neighbor set merging framework (Salehi, Rabiee, & Rajabi, 2012) merges nodes of neighbor set into subgraph by iteration until the condition is not satisfied. But the disadvantage is that it is not suitable for dynamically updated networks, and it has no robustness, and the selection of initial nodes has a great impact on the results. Random walk algorithm (Shenvi, Kempe, & Whaley, 2003) finds some communities by random walk, and the community structures are optimized according to the exchange of nodes between communities. In each iteration, the optimal set of nodes is selected automatically, but it also has high space-time overhead and is not suitable for large-scale networks. In view of the limitations of the above algorithms in mining local network characteristics, a local community detection algorithm based on cluster-first walk and second-cut method (Li, & Pang, 2014) is used to provide important algorithmic support for data analysis in the topic interaction model in MOOC discussion forums in this paper.

Topic Interaction Model for MOOC Discussion Forums

Topic interaction model based on LCD

In order to embody the teaching process of MOOC, the topic interaction model expresses the user's ability to interact in different aspects in the discussion forum as different "documents”. Each "document" divides the user's specific performance ("word") into different levels ("topic"). The model is scored on a four-level system (A-D), its structure diagram is shown in Figure 3.

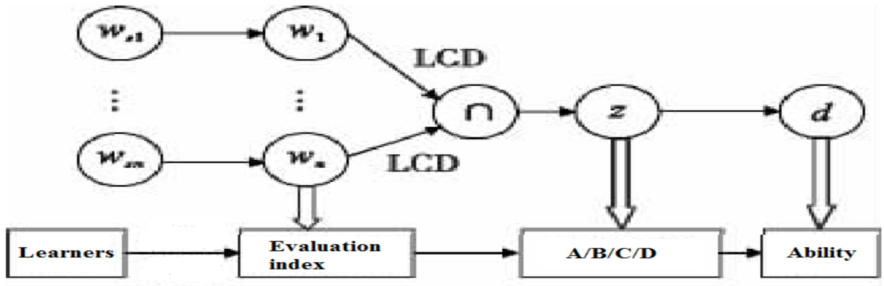

Figure 3. Structure diagram of topic interaction model

Because Gibbs sampling is suitable for classification of documents, and the process of learners' acceptance, understanding and application of new knowledge belongs to subjective activities, it is difficult to determine the classification criteria for specific level, so the concept of local community detection is introduced. The complex network is regarded as a kind of mapping of learner groups, in which learners act as nodes in the network and the interactive relationships between learners act as edges, thus a whole sparse but locally dense topology network is constructed.

Local community detection algorithm uses flow model to treat the network as a power grid， each link is considered as a resistor, and the network topology can be calculated using the voltage or current of the node. Users with typical performance of A, B, C and D levels in *n* specific evaluation aspects $\{w_1, w_2, \dots, w_n\}$, of d_i are regarded as seed nodes:

$$w_{seed} = \{w_{s1}, w_{s2}, \dots, w_{sn}\}$$

$$w_{si} = \{w_{i,A}, w_{i,B}, w_{i,C}, w_{i,D}\}$$

The initial value of the voltage of other nodes except seed node is zero. The voltage of current node *i* is updated by iteration method according to the voltage of adjacent nodes.

$$Vot_i^l = \frac{1}{K_i} \sum_{e_i \in nab(e_i)} Vot_j^{l-1}$$

where K_i represents node degree, l represents iteration times.

The voltage values of all nodes are arranged in descending order, and the Top-N nodes are selected to form the network *G*. In the network G, the voltage of each node is calculated by iteration method, and then they are arranged in descending order to get the ranking result Φ.

For different values of *n*, the Top-n nodes in Φ are taken out as community G'. Community integration $\psi(n)$ are calculated as follows:

$$\psi(\mathrm{n}) = \sum_E \omega_{ij} / \Omega(\mathrm{n})$$

$$\Omega(\mathrm{n}) = \sum_{j}\sum_{i}\omega_{ij}/\Omega(\mathrm{n})$$

where E represents the set of internal edges for the Top-n nodes, ω_{ij} represents edge weight.

For all $\psi(\mathrm{n})$, the peak values are found and the peak set *C* is composed:

$$\psi(\mathrm{n}) = \max(\psi(\mathrm{n}-4{:}\,\mathrm{n}+4))$$

The sharpest peak in set C is found, it is the optimal value:

$$\mathrm{n} = \mathrm{argmax}(\psi(\mathrm{n}-4) + \psi(\mathrm{n}+4) - 2\psi(\mathrm{n}))$$

So the Top-n nodes in Φ are the local community.

As shown in Figure 4, seed nodes are represented by black nodes, for the same level z in d_i, local communities of different *w* overlap with each other. Shadow nodes are located outside the overlapping area, it indicates that some index of the node does not meet the *z* standard. Suppose that the nodes are contained by *t* local communities, if $t \geq n-1$, the node is judged to belong to the level *z*, otherwise it will be automatically reduced to the next level if it does not meet the standard.

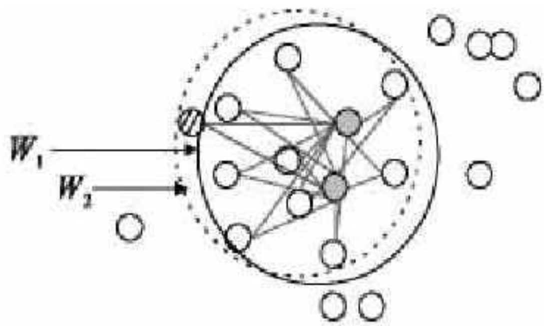

Figure 4. Example of local community detection

Model application in MOOC platform

The course offered by myself on MOOC platform of Chinese University is taken as the research object. The data in the course free discussion forum and the classroom discussion forum are used to classify students' various abilities by the proposed model. In order to accurately evaluate the various capabilities of learners, learners' operation and behavior in teaching and experiment pages are tracked and recorded comprehensively. It is different from the traditional model of "score + final examination + experiment", in the MOOC platform, learners access servers through any computer on the Internet, in the process of watching teaching videos, reading experimental instructions, editing codes and submitting experimental reports, students' online learning efficiency, programming ability and experiment completion are divided into specific evaluation indexes. Specific evaluation indexes are shown in Table 1.

Table 1
Specific Evaluation Indexes of Learners' Performance

d corresponding ability	*w* corresponding evaluation indexes
$\boldsymbol{d_1}$ learning ability	$\boldsymbol{w_1}$ number of repetitions of watching the same video
	$\boldsymbol{w_2}$ time length of fast forward video
	$\boldsymbol{w_3}$ time of reading the lecture notes
	$\boldsymbol{w_4}$ quantity of questions asked
	$\boldsymbol{w_5}$ quality of questions asked
$\boldsymbol{d_2}$ programming ability	$\boldsymbol{w_1}$ average number of code input per second
	$\boldsymbol{w_2}$ programming completion time
	$\boldsymbol{w_3}$ number of times to consult help documents
$\boldsymbol{d_3}$ comprehension ability	$\boldsymbol{w_1}$ number of errors in code running
	$\boldsymbol{w_2}$ experimental data and effects
	$\boldsymbol{w_3}$ completion of the experiment report
	$\boldsymbol{w_4}$ exercise completion time and correct rate

By tracking and collecting the online information, the topic interaction model automatically divides learners into different categories by using LCD algorithm and feeds back to teachers.

Teaching application of topic interaction model

C programming language courses are selected to meet the following requirements: a large number of students, a good reputation of the course, easy to understand the content of the course, and a large number of posts in the discussion forum. Moreover, the selected courses have been completed and the discussion forum has been closed, which makes the collected data objectively, truly and completely reflect the influence of the topic in the discussion process of the discussion forum during the beginning of the course, and the observation variables are simple and controllable.

The course had be held from March 1, 2017 to June 30, 2017. There were 1378 learners selected the course and 7832 posts sent in discussion forum. The collected data were analyzed by factor analysis of variance firstly. The number of learners' responses to participating topics and those of non-participating topics were analyzed to find whether there were significant differences between them.

In order to facilitate the analysis of variance, we evenly selected 50 topics that teachers participated in and 50 topics that teachers did not participate in according to the time distribution. Then the number of student responses to these topics are counted. The number of responses from students on the topic is shown in Table 2.

Table 2
Number of Responses from Students on the Topic

Group	Observation number	Sum	Mean value	Variance
Topics with teachers' participation	50	159	2.9056	7.1652
Topics without teachers 'participation	50	49	0.86468	1.8682

Table 3
Factor Analysis of Variance of Student Responses

	SS	df	MS	F	P-value	F crit
Between groups	116.721	1	116.721	25.078	2.27E-06	3.96265
Within group	486.321	100	4.86321			
Total	603.042	101				

The results of factor analysis of variance of student responses is shown in Table 3.

From Table 3 we can see that there are significant differences (P<0.05) in the number of topic responses by learners in MOOC discussion forum between with teachers ‘participation and those without teachers ‘participation.

Next, the course content is divided into two parts. Some students were taken as experimental subjects to learn the first part of the content, they accept the traditional classroom teaching and the platform teaching respectively. LCD model parameters are set as follows: Iteration times $l = 5$, the maximum number of nodes in a local community (i.e. the maximum number of interactions each time) is 15. After two iterations, truncation is carried out to remove the nodes with less correlation. Then three iterations are performed on the truncated network. The Top-N nodes in Φ were taken to calculate the optimum value of community integration $\psi(n)$. At this time, the N' $(N' \leq N)$ nodes is the student group of corresponding grade *z* and evaluation index *w*.

Ability $d_1 - d_3$ were examined in turn, the intersection of local communities with different evaluation indexes $\boldsymbol{w_1}$-$\boldsymbol{w_4}$ of A-D grade is obtained. Comparing with the reference data, the accuracy of the comprehensive assessment of students’ performance is shown in Table 4.

Table 4
Accuracy of Traditional Assessment and Topic Interaction Model

Ability	Traditional assessment /%	Topic interaction model /%
$\boldsymbol{d_1}$	75.32	89.21
$\boldsymbol{d_2}$	66.12	90.32
$\boldsymbol{d_3}$	78.26	95.78

In traditional teaching, influenced by randomness and various uncertainties, it is impossible to accurately evaluate the teaching effect and learners’ abilities only by the three assessment methods of homework, final examination and experiment. The topic interaction model utilizes various specific evaluation indexes to mine in-depth the valuable information of learners in learning and experimental behavior, it can avoid the greater impact of randomness on the evaluation results, and it can achieve a higher accuracy rate.

On the other hand, based on the topic interaction model, different learning programs with different emphases can be developed for each learner's relatively weak abilities. After discussion and counseling in discussion forum, the learners used the platform to learn the second part of the course.

As shown in Figure 5, 93.8% of the learners’ learning ability, 93.3% of the learners' practical ability and 91.9% of the learners 'understanding ability have been improved.

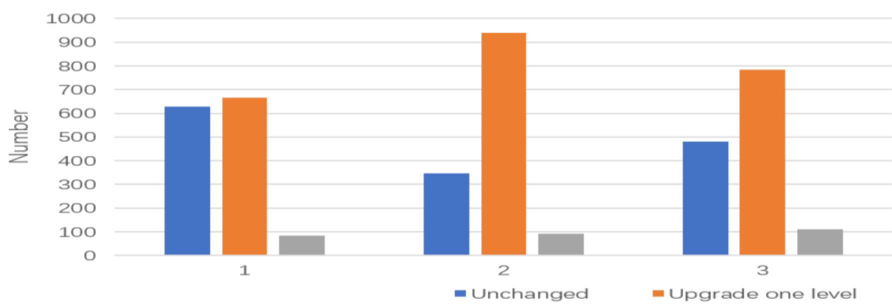

Figure 5. Comparison of evaluation results before and after interaction in forum

Increased learner number of each level after interaction in discussion forums is shown in Figure 6.

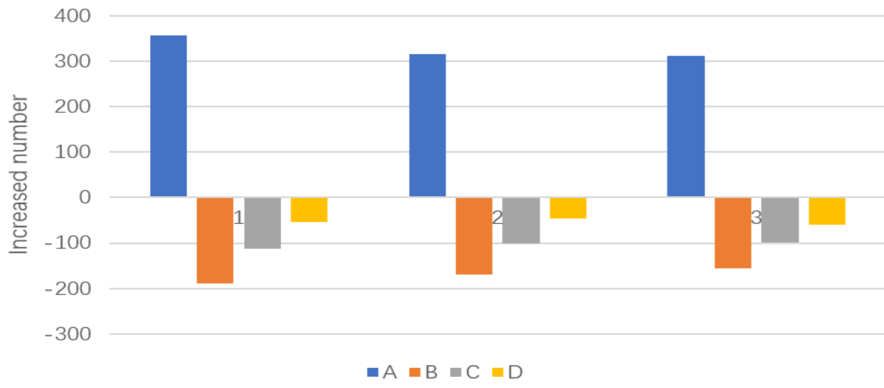

Figure 6. Increased learner number of each level after interaction in discussion forums

From Figure 6 we can see that for ability d_1, d_2 and d_3, 87.2%, 66.9% and 65.7% of the learners whose ability has been upgraded to grade A, respectively.

Conclusion

The proposed topic interaction model in MOOC discussion forum can accurately evaluate learners' learning, programming and understanding ability based on the tracking information of the MOOC platform. The model can be used as an effective basis for teachers to conduct targeted guidance in traditional classroom, it would pay more attention to ability training and comprehensive quality improvement, so it has strong practical value. It can provide a successful precedent and exploration direction for promoting the integration of MOU and higher education in China. How to further improve the accuracy of LCD algorithm and set more comprehensive and effective evaluation indexes is the future research goal.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2012). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3,* 993-1022. http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993

Chen, G., Davis, D., Lin, J., Hauff, C., & Houben, G. J. (2016). Beyond the MOOC platform: Gaining insights about learners from the social web. 15-24. http://dx.doi.org/10.1145/2908131.2908145

Garrison, D. R., Anderson, T., & Archer, W. (1999). Critical inquiry in a text-based environment: Computer conferencing in higher education. *Internet and Higher Education, 2*(2-3), 87-105. http://dx.doi.org/10.1016/s1096-7516(00)00016-6

Gelfand, A. E. (2000). Gibbs sampling. Journal of the American Statistical Association, 95(452), 1300-1304. http://dx.doi.org/10.1080/01621459.2000.10474335

Jian-Hua, Z., & Zheng-You, L. (2014). A sentiment analysis method based on sentiment words extraction and lda feature representation. *Computer and Modernization, 5*, 79-83. http://dx.doi.org/10.3969/j.issn.1006-2475.2014.05.018

Konstan, J. A., Walker, J. D., Brooks, D. C., Brown, K., & Ekstrand, M. D. (2015). Teaching recommender systems at large scale: evaluation and lessons learned from a hybrid MOOC. *ACM Transactions on Computer-Human Interaction (TOCHI), 22*(2), http://dx.doi.org/10.10.1145/2728171

Li, K., & Pang, Y. (2014). A unified community detection algorithm in complex network. *Neurocomputing, 130*(3), 36-43. http://dx.doi.org/10.1016/j.neucom.2012.11.054

Li, Y., Zhou, X., Sun, Y., & Zhang, H. (2016). Design and implementation of Weibo sentiment analysis based on LDA and dependency parsing. *China Communications, 13*(11), 91-105. http://dx.doi.org/10.1109/CC.2016.7781721

Liang, J., Liu, P., Tan, J., & Bai, S. (2014). Sentiment classification based on as-LDA model. *Procedia Computer Science, 31*, 511-516. http://dx.doi.org/10.1016/j.procs.2014.05.296

Ma, J., Zheng, J., & Zhao, G. (2015). The applicable strategy for the courses alliance in regional universities based on MOOC platform. *Procedia-Social and Behavioral Sciences, 176*, 162-166. http://dx.doi.org/10.1016/j.sbspro.2015.01.457

Moctar, A. O. M., Sarr, I. (2016). Static and dynamic community detection. *Revue d'Intelligence Artificielle, 30*(4), 469-496. http://dx.doi.org/10.3166/RIA.30.469-496

Mothe, J., Rakotonirina, A. J. (2018). Contextual collaborative filtering. A LDA-based approach. *Ingenierie des Systemes d'Information*, *23*(1), 89-109. http://dx.doi.org/10.3166/ISI.23.1.89-109

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature, 435*(7043), 814-814. http://dx.doi.org/10.1038/nature03607

Salehi, M., Rabiee, H. R., & Rajabi, A. (2012). Sampling from complex networks with high community structures. *Chaos, 22*(2), 2202-2229. http://dx.doi.org/10.1063/1.4712602

Seth, S., Bhattacharyya, D., & Kim, T. H. (2014). CBACCN: Constraint based community discovery in complex networks. *International Journal of Applied Engineering Research*, *9*(23), 18115-18127.

Shenvi, N., Kempe, J., & Whaley, K. B. (2003). Quantum random-walk search algorithm. *Physical Review A, 67*(5), 125-128. http://dx.doi.org/10.1103/physreva.67.052307

Thakur, G. S., Tiwari, R., Thai, M. T., Chen, S. S., & Dress, A. W. M. (2009). Detection of local community structures in complex dynamic networks with random walks. *IET Systems Biology, 3*(4), 266-270. http://dx.doi.org/10.1049/iet-syb.2007.0061