



Ranking Research Institutions Based on the Combination of Individual and Network Features

Wenyan Wang¹, Guangyu Wang¹, Jun Zhang², Peng Chen²,
and Bing Wang^{1,2}(✉)

¹ School of Electrical and Information Engineering,
Anhui University of Technology, Maanshan 243002, People's Republic of China
wangbing@ustc.edu

² The Institute of Health Sciences, Anhui University, Hefei 230601,
Anhui, China

Abstract. Regarding the fierce competition between research institutions, institutional rankings are widely carried out. At present, there are many factors affecting the ranking of institutions, but most of them are aimed at the attributes of the institutions themselves, and the feature selection is relatively simple. Therefore, this paper proposes a state-of-the-art method combining different types of features for predicting the influence of scientific research institutions. Based on the MAG dataset, this paper first calculates the institutional scores through the publication volume of the article, constructs an inter-institutional cooperation network, and calculates the importance characteristics of the institutions in the network. Then, considering the contribution of the faculty and staff to the organization, an individual characteristic based on the author's influence is constructed. Finally, a random forest algorithm is used to solve this prediction problem. As a result, this paper raises the ranking accuracy rate NDCG@20 to 0.865, which is superior to other methods. The experimental results show that this method has a good effect on the prediction of innovation capability.

Keywords: Institutional ranking · Feature combination · MAG · Random forest

1 Introduction

In recent years, the activities of ranking research institutes have flourished and developed, which is the product of the development of research institutes to a certain stage [1, 2]. However, the development of ranking research institutions is not perfect since most ranking methods are static and using only part of indicators. The KDD Cup 2016 would like to galvanize the community to address this very important problem through any publicly available datasets, like the Microsoft Academic Graph (MAG). The Microsoft Academic Graph is a heterogeneous graph containing scientific publication records, citation relationships between those publications, as well as authors, institutions, journals, and fields of study [3].

For institutional rankings, good features is the key to reliable ranking results. Currently, there are about three major types of features were established. Firstly, based on institution attributes, Gupta et al. only use the feature of the affiliation score calculated from the amount of article received at a meeting to make predictions [6, 7]. Wilson et al. firstly classified the papers to determine whether the paper belonged to the full research papers or to all papers. The motivation for this is that some conferences listed by kdd cup 2016 receive only the full research papers [8–10]. In addition, the total number of papers of each institution is also used as a unique feature that replaces the affiliation score [9, 11]. Similarly, the inherent property of an institution, such as name and location of affiliations, state GDP, etc., can also be used as features of the institution. Last but not least, the relations between institutions were also taken into consideration to represent the importance of institutions [7, 10–12]. Secondly, paper features were established to demonstrate the paper’s influence on the institution. Such as the number of papers, the trend of paper number and track record describing the publication history of an institution, and so on [4, 5, 8]. Lastly, the individual characteristics of the author had also been excavated. Such as, the feature of active degree measuring the institution active degree in one conference according to the number of active authors in the institution, the continuity evaluate feature representing how the research filed of an institution is insisting on were designed [5], and number of first, second authors and (author, paper) pair can also make a contribution to the assessment of institutional influence. What’s more, the network of authors is also established to highlight the more influential authors [13–15]. In fact, in order to expand the limited availability of data sets, information about some related meetings can also be grouped together as a common feature [12, 13, 15].

In this work, although the scores of each institutions from 2011 to 2015 can be calculated directly from the dataset provided by the KDD Cup, it is not ideal to predict the institution score of 2016 by only using this single feature. On the issue of extreme lack of institutional information, it is urgent to construct more characteristics that describe the attributes of institutions. Therefore, in this paper, the individual characteristics based on the author and the network characteristics based on the institution were constructed from the micro and macro perspectives, and these features were treated as datasets and applied to random forest models to solve institutional ranking problems. As a result, our work raises the ranking accuracy NDCG@20 to 0.865, which outperformed current methods and the complementary information between the features had been proved.

2 Materials and Feature Generation

2.1 Dataset

Although the organizers of KDD Cup encourage the participants to use any publicly available information, they do provide us with Microsoft Academic Graph (MAG). The data can be downloaded from the website of <http://aka.ms/academicgraph> and the version “2016-02-05” is available. All the data is uniquely identified by the primary

key, such as the conference ID, affiliation ID, etc. Take the year of 2015 as an example, some of the data in the MAG dataset shown in Table 1.

Table 1. Examples of MAG dataset.

Paper	Year	Author
76381EFA	2015	8084BB24
76381EFA	2015	75421677
76381EFA	2015	832818A2
716E3093	2015	10C312AC
8389DBD7	2015	112ACB8A

2.2 Feature Generation

2.2.1 Affiliation Score

According to the cooperation among authors, institutions and papers displayed in dataset, as shown in Table 1, and following the simple policy specified by organizers. The most direct and effective affiliation score can be calculated firstly since it has a fundamental effect on the predicting institutions future scores and it’s the easiest to build in an existing data set. Figure 1 shows the method of calculating the affiliations score.

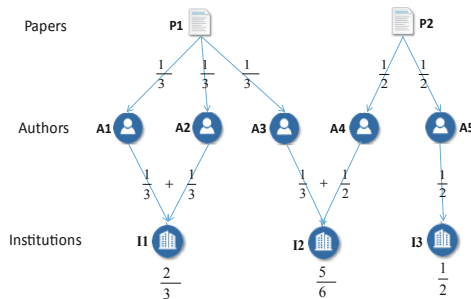


Fig. 1. Example of institutional score calculation method

2.2.2 Network Features

After statistics on the numbers of articles that belonging to one or more institutions published in the KDD conference from 2011 to 2015, as shown in Table 2, we found that articles written by multiple affiliations accounted for a large proportion. So we take the year of 2015 as an example, constructing the co-operative networks and the node degree distribution histogram, as shown in Fig. 2. In the network, each node (red circle) represents an affiliation, and they are connected by a straight line.

From the Fig. 2(a), we can see that some nodes occupy the central position in the network and have obvious importance. And in Fig. 2(b), it shows that the distribution of the node degree satisfies the power law distribution so that the graph is a scale-free

network [16–18]. Based on these two factors, the collaboration networks from 2011 to 2015 were established.

Table 2. The Number of papers in KDD.

Year	Multiple affiliation	Single affiliation	Total
2011	71	79	150
2012	73	58	131
2013	67	58	125
2014	79	68	147
2015	100	60	160

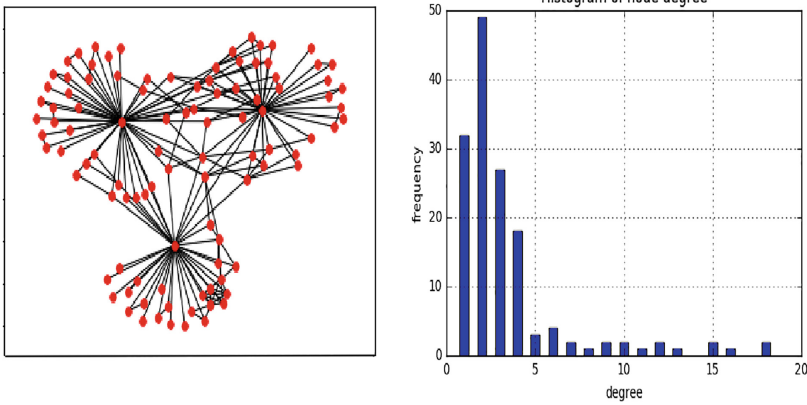


Fig. 2. (a) Affiliation collaboration network (b) Histogram of node degree (Color figure online)

In the analysis of collaboration network, degree centrality (DC), closeness centrality (CC) and betweenness centrality (BC) are the ways to measure the importance of nodes in a network. Therefore, this paper use them to evaluate the institutional innovation capability [19].

Degree centrality is defined as the number of edges incident upon a node. In order to make a comparison between different scale networks, the degree centrality of node i was calculated using (1). Closeness centrality reflects the degree of the node to the network center in the whole network structure. Using (2), the closeness centrality of node i was calculated. And betweenness centrality is defined as the more number of shortest paths through a node, the greater important of this node in the shortest path of all node pairs. Using (3), the betweenness centrality of node i was calculated.

$$DC(i) = \frac{k_i}{N - 1} \tag{1}$$

$$CC(i) = \frac{N - 1}{\sum_{j \neq i} d_{ij}} \tag{2}$$

$$BC(i) = \sum_{s \neq i \neq t} g_{st}^i \tag{3}$$

Where $k_i = \sum_{i=1}^N a_{ij}$, N is the number of all nodes, $N - 1$ is the maximum possible edges value of the node, and a_{ij} is an element with i row j column of the adjacency matrix, d_{ij} is the shortest path between node i and j . g_{st} is the total number of shortest path from node s to node t , and g_{st}^i is the number of shortest path through node i in the shortest path from node s to t [19, 20].

2.2.3 Individual Features of the Author

Inspired by the cooperation between affiliations, author collaboration network has also been established. In order to highlight the importance of the author, the number of cooperations between authors is used as the weight of the connection line.

In the author’s collaboration network, this paper first assumes that the author’s importance is only related to the authors in the two nearest layers of the network. For each author, the collaboration score can be expressed as

$$ACS(A_i) = \frac{N_1}{N_2} * (\sum_{j \in layer1} W_j + \sum_{j \in layer2} W_k) \tag{4}$$

Where $ACS(A_i)$ is the author collaboration score of A_i , N_1 and N_2 are the number of authors included in the first and second layers, respectively. W_j , W_k are the weights corresponding to the author in the first and second layers.

The probability of an author publishing paper (PAPP) in the year $(t + 1)$ is computed by taking the ratio of total number of years the author has published papers in the past to total number years under consideration (t) . The probability score thus calculated is a binomial probability for publishing paper in year $(t + 1)$.

$$P_{a,t+1} = \frac{\sum_{i=1}^t x_{a,i}, \text{ for } x_{a,i} = 1}{t} \tag{5}$$

The Joint probability 1 (JP1) for publishing paper in year (t) and in the year $(t + 1)$ is computed by taking the ratio of the frequency of paper published in consecutive years to the sum of frequencies of publishing paper in year (t) and year $(t + 1)$ plus publishing paper in year (t) but not in $(t + 1)$.

$$p(x_{a,t+1} = 1, x_{a,t} = 1) = \frac{\sum_{i=1}^t x_{a,i}, \text{ for } x_{a,i} = 1 \text{ and } x_{a,i} = 1}{(\sum_{i=1}^t x_{a,i}, \text{ for } x_{a,i} = 1 \text{ and } x_{a,i-1} = 1) + (\sum_{i=1}^t x_{a,i}, \text{ for } x_{a,i} = 0 \text{ and } x_{a,i-1} = 1)} \tag{6}$$

The Joint probability 2 (JP2) for not publishing paper during the year (t) but publishing paper in the year ($t + 1$) is computed by taking the ratio of the frequency of not publishing paper in the year (t) and publishing paper in the year ($t + 1$) to the sum of frequencies of publishing paper in year (t) and year ($t + 1$) plus not publishing paper in year (t) but in ($t + 1$).

$$P(x_{a,t+1} = 1, x_{a,t} = 0) = \frac{\sum_{i=1}^t x_{a,i} = 1, \text{for } x_{a,i} = 1 \text{ and } x_{a,i-1} = 0}{\left\{ \left(\sum_{i=1}^t x_{a,i}, \text{for } x_{a,i} = 1 \text{ and } x_{a,i-1} = 1 \right) + \left(\sum_{i=1}^t x_{a,i}, \text{for } x_{a,i} = 1 \text{ and } x_{a,i-1} = 0 \right) \right\}} \tag{7}$$

Where $x_{a,i} = 1$ if the author published a paper in that year, else $x_{a,i} = 0$, t represents a time range under consideration.

3 Individual Features of the Author

3.1 Data Normalization

In this work, each sample was represented by three type of features. However, these features contain different physical meanings or data ranges. So, in order to eliminating the impact of imbalanced information expression, all of features have to be normalized before applying to the model. Here, all values of each feature always fall within a fixed interval $[0, 1]$ by

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{8}$$

Where x_{\max}, x_{\min} represent the maximum, minimum of variables, respectively.

3.2 Random Forest Regression Model

Random forest uses the bootstrap resampling method to extract multiple samples from the original sample and building a model of decision tree for each bootstrap sample. This method was proved to have higher prediction accuracy, good tolerance to outliers and noise and it is not easy to over-fitting [20, 21].

The fitting process of a single decision tree in the random forest algorithm is as follows:

- (a) Bagging sampling is used to form the training set that equal to the original sample number.
- (b) The feature is chosen randomly when the internal node splits.
- (c) Don't prune every tree in the forest, let it grow randomly [22].

Suppose the input space is divided into M units R_1, R_2, \dots, R_m , and each unit R_m has a fixed output value c_m , the model of regression tree can be expressed as below:

$$f(x) = \sum_{m=1}^M c_m I(x \in R_m) \tag{9}$$

In this paper, through the grid search algorithm, the optimal number of trees and the ratio between the features required for each tree and the total features are found, 150 and 0.7 respectively, and there has not been a fitting phenomenon, so the pruning operation is not used in this work.

3.3 Evaluation Strategy

Normalized Discounted Cumulative Gain (NDCG) is one of the most popular ranking indicators. According to the requirements of the organizers, we only need to focus on the top 20 institutions. On the basis of this notion, NDCG@20 was adapted to measure the relevance and it can be computed as follows [23]

$$DCG@20 = \sum_{i=1}^{20} \frac{rel_i}{\log_2(1+i)} \tag{10}$$

$$NDCG@20 = \frac{DCG@20}{Ideal\ DCG@20}$$

Where i is the rank of an institution, and rel_i is this institution’s relevance score.

4 Experimental Results and Discussion

4.1 Features and Samples Selection

For features, affiliation score are considered to be inherent features and they are combined with individual and network features respectively as authors’ individual feature and institutional centrality feature. At the same time, the individual and network features are combined, called the fusion feature. Finally, three kinds of feature sets were constructed as shown in Table 3.

Table 3. Features in training and testing set.

Features				Time intervals	
Fusion features	Institutional centrality feature	Network centrality	BC	2011–2015	
			CC	2011–2015	
			DC	2011–2015	
	Individual feature	Affiliation score	Authors’ individual	AS	2011–2015
				ACS	2011–2015
				JP1	2014–2015
				JP2	2014–2015
				PAPP	2014–2015

4.2 Results

Although this paper generates a variety of features, it is not clear whether they are really related to the influence of the organization. To verify this problem, the correlation coefficients between each feature and the corresponding institutional score (AS) for 2011-2015 were calculated and shown in the Table 4.

Table 4. The correlation coefficient between each feature and the corresponding institutional score

Year	BC	CC	DC	JP2	JP1	ACS	PAPP
2011	0.71	0.49	0.73			0.96	
2012	0.66	0.56	0.75			0.91	
2013	0.64	0.57	0.59			0.91	
2014	0.65	0.6	0.64	0.93	0.68	0.84	0.94
2015	0.75	0.47	0.74	0.94	0.76	0.91	0.95

From the data in Table 4, the correlation coefficient between the closeness centrality (CC) and the institutional scores in 2011-2015 are less than or equal to 0.6. The correlation coefficients of ACS were distributed around 0.9, which was significantly higher than the closeness centrality. In order to obtain the effect of each feature and its correlation coefficient on the forecasting influence of the institutions, this paper first uses the network features and fusion features including closeness centrality and not including closeness centrality to predict the institutional influence, respectively. The violin chart of 10 prediction results and average values are shown in Fig. 3(a). In addition, this paper attempts to make the effect of the betweenness centrality on the predictive mechanism's influence. The result is shown in Fig. 3(b). Where fusion feature-CC represents the prediction result after the closeness centrality is removed, fusion feature-BC represents the result after the closeness centrality is removed and betweenness centrality is also removed, and InCe in Fig. 3(b) represents the institution centrality feature. It can be seen from Fig. 3 that after removing the closeness central characteristics, the prediction accuracy and robustness of using only the institution centrality feature are improved, and the performance of using the fusion feature prediction is also improved. From Fig. 3(a), it can be seen that the average accuracy of the institution centrality and fusion feature is increased by about 3%, the height of the fusion feature-CC in the violin diagram is significantly lower, and the In Ce-CC is also shown the same performance in the violin diagram of the Fig. 3(b), which indicates the closeness centrality established in this paper does not apply to institutional impact prediction. However, from Fig. 3(b), it can be found that the removal of the betweenness centrality is the opposite of the removal of the closeness centrality. Although the stability of the fusion feature is increased, the prediction accuracy of the institution centrality and the fusion feature are reduced by 4% and 1% respectively.

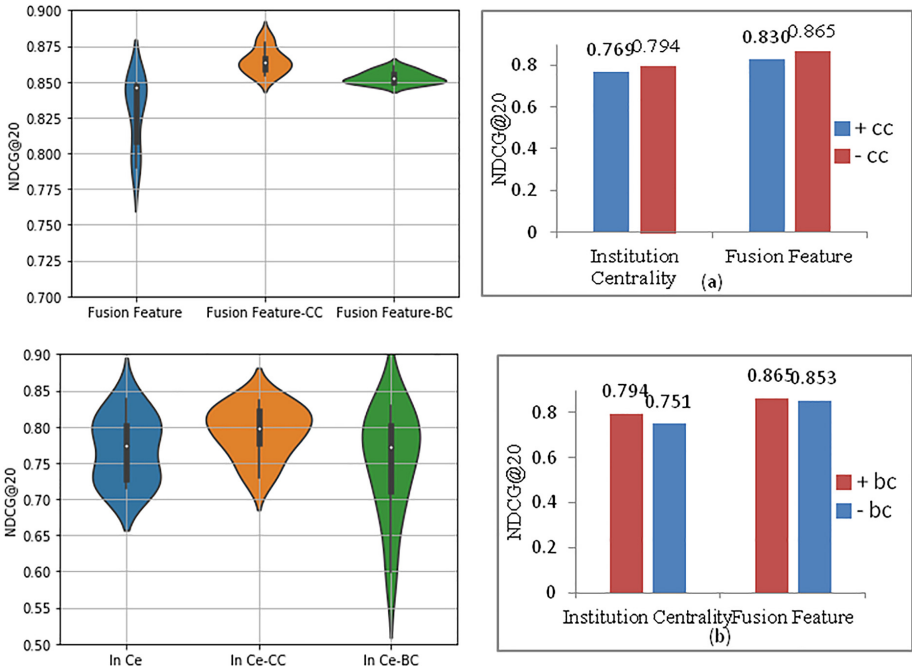


Fig. 3. Prediction results using different time periods.

In Table 4, after processing the features longitudinally with correlation coefficient values, this paper sorts all the generated features by the value of the feature correlation coefficients, and eliminates the features with correlation coefficients less than 0.6, 0.7, 0.8, and 0.9. The remaining features are used for prediction to observe the influence of each feature in the horizontal direction on the ranking prediction. The box plot of the predicted result is shown in Fig. 4. In the box plot, a line in the middle of the box indicates the median of the data. The upper and lower lines of the box indicate the upper quartile and the lower quartile of the data. A line above and below the box represents the maximal and minimum values, the circle represents the outlier. It can be seen from Fig. 3 and Table 4 that the smaller the feature correlation coefficient of the removal, the larger the prediction accuracy and the more stable. In other words, the more the feature amount used, the higher the prediction accuracy, the more robust. This further proves that there is information complementation between the features constructed in this paper, and the fusion features have the best predictive performance. Therefore, this paper will use all the features except the closeness centrality to predict the influence, and the prediction accuracy is 0.865.

Finally, the central, individual and fusion features of the institutions are respectively applied to the random forest model, and the predicted results are shown in Fig. 5. In order to prove the validity of the confused features in this paper, we compare the prediction results with those who did the same work as described in the introduction.

The used features and prediction results are shown in Table 5. Experiments result proves that the fusion of these two features in this paper is more effective in predicting institutional influence.

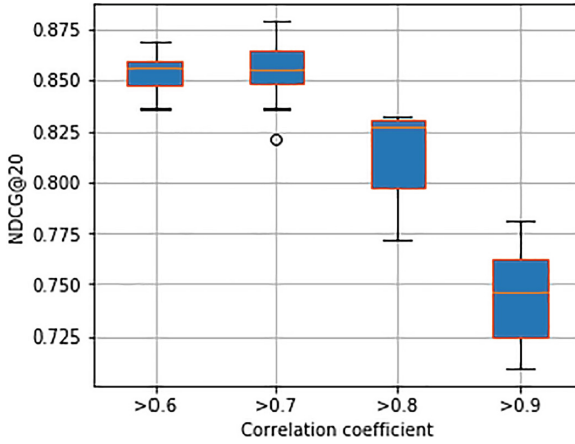


Fig. 4. Prediction results with different correlation coefficients.

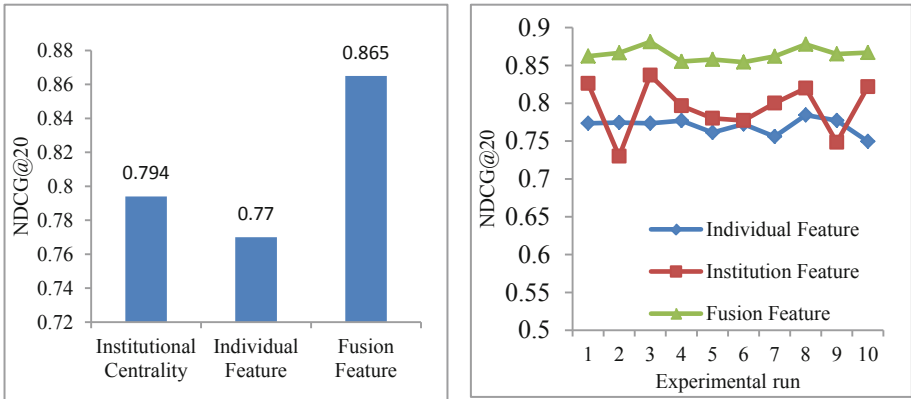


Fig. 5. Prediction results with three types of features.

Table 5. Comparison of results.

Method	Features	Prediction result
Paper 1 [11]	Network central features + Human knowledge	0.794
Paper 2 [14]	Individual features	0.77
This paper	Network central feature + Individual feature	0.865

5 Conclusion

In order to improve the prediction accuracy of scientific research institutions' innovative ability, a novel combined the individual and network attributes method had been proposed in this paper, which is based on the random forest algorithm. By analyzing the correlation coefficient between the constructed feature and the institutional score, As a result, this paper uses all features except closeness centrality, and the central, individual and fusion features of the institutions are respectively applied to the random forest model to realize the prediction of innovation ability of scientific research institutions. Experimental results demonstrate that the feature generate method proposed in this paper has a great effectiveness in both accuracy and stability of the institution influence prediction, and the information between the features is complementary.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (Nos. 61472282, 61672035, and 61872004), Anhui Province Funds for Excellent Youth Scholars in Colleges (gxyqZD2016068), the fund of Co-Innovation Center for Information Supply & Assurance Technology in AHU (ADXXBZ201705), and Anhui Scientific Research Foundation for Returned Scholars.

References

1. Cuthbert, R.: University rankings, diversity, and the new landscape of higher education. *Int. J. Lifelong Educ.* **30**, 119–121 (2011)
2. Szentirmai, L., Radacs, L.: World university rankings qualify teaching and primarily research. In: *IEEE International Conference on Emerging Elearning Technologies and Applications*, pp. 369–374 (2013)
3. Sinha, A., et al.: An overview of microsoft academic service (MAS) and applications. In: *International Conference on World Wide Web*, pp. 243–246 (2015)
4. Mussard, M., James, A.P.: Engineering the global university rankings: gold standards, its limitations and implications. *IEEE Access* **PP**, 1 (2018)
5. Al-Juboori, A.F.M.A., Su, D.J., Ko, F.: University ranking and evaluation: trend and existing approaches. In: *The International Conference on Next Generation Information Technology*, pp. 137–142 (2011)
6. Gupta, A., Murty, M.N.: Finding influential institutions in bibliographic information networks (2016)
7. Orouskhani, Y., Tavabi, L.: Ranking research institutions based on related academic conferences. *arXiv e-prints* (2016)
8. Wilson, J., Mohan, R., Arif, M., Chaudhury, S., Lall, B.: Ranking academic institutions on potential paper acceptance in upcoming conferences (2016)
9. Sandulescu, V., Chiru, M.: Predicting the future relevance of research institutions - the winning solution of the KDD Cup 2016 (2016)
10. Zhang, J., Xu, B., Liu, J., Tolba, A., Al-Makhadmeh, Z., Xia, F.: PePSI: Personalized prediction of scholars' impact in heterogeneous temporal academic networks (2018)
11. Klimek, P.S., Jovanovic, A., Eglhoff, R., Schneider, R.J.S.: Successful fish go with the flow: citation impact prediction based on centrality measures for term–document networks. *Scientometrics* **107**(3), 1265–1282 (2016)

12. Xie, J.: Predicting institution-level paper acceptance at conferences: a time-series regression approach (2016)
13. Qian, Y., Dong, Y., Ma, Y., Jin, H., Li, J.: Feature engineering and ensemble modeling for paper acceptance rank prediction (2016)
14. Moed, H.: Bibliometric rankings of world universities (2006)
15. Bai, X., Zhang, F., Hou, J., Xia, F., Tolba, A., Elashkar, E.: Implicit multi-feature learning for dynamic time series prediction of the impact of institutions. *IEEE Access* **PP**, 1 (2017)
16. Crucitti, P., Latora, V., Marchiori, M., Rapisarda, A.: Error and attack tolerance of complex networks. *Nature* **340**, 378–382 (2000)
17. Holme, P., Edling, C.R., Liljeros, F.: Structure and time evolution of an Internet dating community. *Soc. Netw.* **26**, 155–174 (2004)
18. Barabási, A.L., Jeong, H., Nédá, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. *Physica A Stat. Mech. Appl.* **311**, 590–614 (2002)
19. Ren, X., Lü, L.: Review of ranking nodes in complex networks. *Chin. Sci. Bull.* **59**, 1175 (2014)
20. Belgiu, M., Drăguț, L.: Random forest in remote sensing: a review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **114**, 24–31 (2016)
21. Désir, C., Bernard, S., Petitjean, C., Heutte, L.: One class random forests. *Pattern Recogn.* **46**, 3490–3506 (2013)
22. Zhou, Z.H.: Ensemble learning. In: *Encyclopedia of Biometrics*, pp. 270–273 (2009)
23. Wang, Y., Wang, L., Li, Y., He, D., Liu, T.Y., Chen, W.: A theoretical analysis of NDCG type ranking measures. *J. Mach. Learn. Res.* **30**, 25–54 (2013)