Topic based Research Competitiveness Evaluation

Yue Mingliang[1] Ma Tingcan[2*]

[1] *yueml@whlib.ac.cn* [2] *matc@whlib.ac.cn*
Wuhan Documentation and Information Center, Chinese Academy of Sciences, Wuhan (China)

Abstract
Research competitiveness analysis refers to the measurement, comparison and analysis of the research status (i.e., strength and/or weakness) of different scientific research bodies (e.g., institutions, researchers, etc.) on different research fields. Improving research competitiveness analysis method can be conducive to accurately obtaining the research status of research fields and research bodies. This paper presents a method of evaluating the competitiveness of research institutions based on research topic distribution. The method uses the LDA topic model to obtain paper-topic distribution matrix to objectively assign the academic impact of papers (such as times of citation) to research topics. Then the method calculates the competitiveness of each research institution on each research topic with the help of institution-paper matrix. Finally, the competitiveness and the research strength and/or weakness of the institutions are defined and characterized. Case study shows that the method can lead to an objective and effective evaluation of the research competitiveness of given research institutions on given research field.

Conference Topic
The theory, method and principle of five metrics science concepts, that is, Bibliometrics, Informetrics, Scientometrics, Webometrics and Knowledgometrics.

Introduction
Research competitiveness analysis refers to the measurement, comparison and analysis of the research status of different scientific research bodies (e.g., institutions, researchers, etc.) on different research fields (Zhang, 2014). Improving competitiveness analysis methods can be conducive to obtaining the research status of research bodies, clarifying their strengths and/or weaknesses, and in turn promoting collaborative innovation among different research bodies and different research fields.

Generally, research competitiveness analysis is carried out based on research papers and involves three steps, i.e., research field (topic) recognition, competitiveness evaluation and competitiveness analysis (Gei, 2013). Research field recognition is important since different research fields are always not comparable, while topic recognition can result in a fine granular evaluation for strength and/or weakness characterization. For field recognition, a paper's research filed is usually determined based on partition standard provided by the scientific literature database providers, e.g., ESI, Incites, Wos, etc. (Chen & Shi, 2013; Dong, 2014; Li 2012; Cova, 2013). For research topic, it is often represented by keywords with high frequencies and their frequent combinations derived using certain analysis tools (e.g., CiteSpace) (Chen, 2006; Chen & Hu, 2012). When the fields (topics) are determined, the scientometrics criterions of papers are used for competitiveness evaluation of the corresponding institutions on the corresponding research fields (topics). Those criterions may include paper count, paper IFs, paper citation counts, etc. (Mkhnacheva, 2011; Morris, 2003; Small, 2009; Shibata, 2008). Finally, the ranking of competitiveness is given and the strengths and/or weaknesses are analysed (Liu, 2015; Small, 2014).

Those methods have made concrete progress on competitiveness evaluation; however, the following problems should be further considered. First, the mapping from papers to research fields and topics are too straightforward for a precise evaluation, since many multi-discipline papers cannot be simply partitioned into a unique research field, and a small set of frequent

* Corresponding Author

keywords may not be capable to represent a research topic. Second, in the current works, papers are all bounded to a unique research field, and papers relating to multiple research topics are considered contributing equally to each of the topics. However, this may not always intuitive since a paper always has certain main research points corresponding to one or more (but not all the relating) field(s) and/or topic(s).

Focusing on the mentioned problems, this paper presents a method of evaluating the competitiveness of research institutions based on research topic distribution. The method uses the LDA topic model to obtain paper-topic distribution matrix to objectively assign the academic impact of papers (such as the number of cited times) to research topics. Then the method calculates the impact of each research institution on each research topic with the help of institution-paper matrix. Further, the competitiveness scores of institutions are calculated and the research strength and/or weakness of the institutions are defined and derived. Finally, case study is carried out to show effectiveness of the proposed evaluation method. It is to be noted that in the proposed method, there is no need to distinguish research field and topic, since a research field can be viewed as a higher abstraction of research topics. That means by setting proper parameters, LDA can be used to model paper-field distribution.

Evaluation Method

The proposed method goes through the following steps for evaluation: 1) topic recognition, 2) impact allocation, 3) competitiveness measurement. We explained each step as follows.

Topic recognition

LDA is a document topic generation model (Blei, 2003). The model presumes that the words in the topic and the topics of the document are both subject to certain polynomial distributions. Hence generating a document can be seen as a repeated process of selecting a topic with a certain probability and then selecting a word in the topic with a certain probability. The model can be formally represented as $\Omega = \Phi \times \Theta$, where Ω, Φ and Θ is document-word distribution, topic-word distribution and document-topic distribution respectively, $\times$ represents matrix multiplication, as demonstrated in Fig. 1. In Fig 1, we have 3 papers; each is composed of 2 topics and 3 words. The LDA model can be used to determine Θ for a set of documents by setting a proper topic number *n*.

$$\begin{matrix} & \Omega & & & \Phi & & & \Theta & \\ \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} \\ \omega_{21} & \omega_{22} & \omega_{23} \\ \omega_{31} & \omega_{32} & \omega_{33} \end{bmatrix} & = & \begin{bmatrix} \varphi_{11} & \varphi_{12} \\ \varphi_{21} & \varphi_{22} \\ \varphi_{31} & \varphi_{32} \end{bmatrix} & \times & \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \end{matrix}$$

Figure 1 LDA Model of 3 papers, 2 topics and 3 words

Impact allocation

Once we got the paper-topic distribution matrix Θ, we can then allocate the academic impact of the papers to the relating topics using values in Θ as weights to get paper-impact matrix Γ (that characterizes the impact of papers on topics). More formally, given *m* papers, *n* topics, paper-topic distribution matrix $\Theta = \{\theta_{ij} \mid \theta_{ij} \in [0,1], i \in [1,m], j \in [1,n], \sum_j \theta_{ij} = 1\}$, paper-impact vector $\mathrm{I} = < \iota_1, \iota_2, \ldots, \iota_m >^T$, the paper-impact matrix Γ can be calculated as $\Gamma = \Theta^T \otimes \mathrm{I} = \{\gamma_{ij} \mid \gamma_{ij} = \theta_{ij} \times \iota_i, \theta_{ij} \in \Theta^T, \iota_i \in \mathrm{I}\}$, where θ_{ij} is the weight of paper *i* on topic *j*, ι_i is the impact indicator of paper *i*, γ_{ij} is the calculated impact value of paper *i* on topic *j*. Fig. 2 (a) shows an example with 3 papers and 2 topics.

$$\underset{\Gamma}{\begin{bmatrix} 0.26 & 0.26 \\ 0.252 & 0.588 \\ 0.4 & 0.1 \end{bmatrix}} = \underset{\Theta^T}{\begin{bmatrix} 0.5 & 0.5 \\ 0.3 & 0.7 \\ 0.8 & 0.2 \end{bmatrix}} \otimes \underset{\mathrm{I}}{\begin{bmatrix} 0.52 \\ 0.84 \\ 0.50 \end{bmatrix}} \qquad \underset{\mathrm{I}}{\begin{bmatrix} 0.52 \\ 0.84 \\ 0.50 \end{bmatrix}} = \underset{\Lambda}{\begin{bmatrix} 0.2 & 1 \\ 1 & 0.6 \\ 0.7 & 0.2 \end{bmatrix}} \times \underset{\mathrm{A}}{\begin{bmatrix} 0.6 \\ 0.4 \end{bmatrix}}$$

(*a*) (*b*)

Figure 2 Example of (*a*) calculation of paper-impact matrix and (*b*) calculation of paper-impact vector

In practice, a paper's academic impact may relate to many aspects, e.g., times of citation, impact factor, and etc. We use a weighted composition of the aspects to make an overall evaluation of paper impact. That is, suppose we have *l* factors (that influence academic impact of a paper) whose values are given in the paper-impact matrix $\Lambda = \{\lambda_{ij} \mid i \in [1,n],\ j \in [1,l]\}$, the paper-impact vector is calculated as $\mathrm{I} = \Lambda \times \mathrm{A}$, where λ_{ij} is the normalized impact value of paper *i* on factor *j*, $\mathrm{A} = <\alpha_1, \alpha_2, \ldots \alpha_l>^T$ gives the weights determining the preferences of every factor during composition. Fig. 2 (b) demonstrates an example of compositing 2 factors. After the paper-impact matrix Γ is obtained, we can now characterize the impact of institutions on various topics. Suppose we have *v* institutions, given institution-paper matrix $\Psi = \{\psi_{ij} \mid \psi_{ij} \in \{0,1\}, i \in [1,m],\ j \in [1,v]\}$, the institution-impact matrix Ξ can be calculated as $\Xi = \Psi \times \Gamma$, where $\psi_{ij} = 1$ means institution *i* has authorship with paper *j*, $\psi_{ij} = 1$ means the opposite, each element $\xi_{ij} \in \Xi$ is the calculated impact value of institution *i* on topic *j*. Fig. 3 presents an example of the calculation of Ξ of 3 institutions.

$$\underset{\Xi}{\begin{bmatrix} 0.652 & 0.688 \\ 0.512 & 0.848 \\ 0.252 & 0.588 \end{bmatrix}} = \underset{\Psi}{\begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}} \times \underset{\Gamma}{\begin{bmatrix} 0.26 & 0.26 \\ 0.252 & 0.588 \\ 0.4 & 0.1 \end{bmatrix}}$$

Figure 3 Calculation of institution-impact matrix

Competitiveness measurement

Now we can measure the competitiveness of the relating institutions based on institution-impact matrix Ξ. First, we calculate the average impact of all the institutions on each topic: $\xi = <\xi_1, \xi_2, \ldots, \xi_n>$, where $\forall j \in [1,n], \xi_j = \sum_{i=1}^{v} \xi_{ij} / v, \xi_{ij} \in \Xi$. Then, we calculate the difference matrix Δ between Ξ and ξ as $\Delta = \Xi \ominus \xi = \{\delta_{ij} \mid \delta_{ij} = \xi_{ij} - \xi_j, \xi_{ij} \in \Xi, \xi_j \in \xi\}$, as exemplified in Fig. 4.

$$\underset{\Delta}{\begin{bmatrix} 0.18 & -0.02 \\ 0.04 & 0.14 \\ -0.22 & -0.12 \end{bmatrix}} = \underset{\Xi}{\begin{bmatrix} 0.652 & 0.688 \\ 0.512 & 0.848 \\ 0.252 & 0.588 \end{bmatrix}} \ominus \underset{\xi}{\begin{bmatrix} 0.472 & 0.708 \end{bmatrix}}$$

Figure 4 Calculation of difference matrix

The difference matrix Δ can then be used to assess the research status of the institutions. Given a threshold τ and a percentage *p*, we can define, for example, 1) if $\delta_{ij} \geq \tau$, then

research topic j is the strength of institution i; 2) if $\delta_{ij} \le -\tau$, then research topic j is the weakness of institution i; for a certain topic j, after ranking the institutions according to δ_{ij}, let R_{ij} denote the ranking of institution i on topic j, 3) if $R_{ij} \le pv$, then institution i is leading the research on topic j, and many other rules. Based on the definition, as to our example in Fig. 4, if we set $\tau = 0.1$, then we can say that topic 1 and 2 is the strength of institution 1 and 2 respectively; both topic 1 and 2 are the weaknesses of institution 3.

Case Study

In this section, we use case study to verify the effectiveness of the proposed evaluation method from two perspectives: 1) the rationality of LDA model, 2) the effectiveness of competitiveness evaluation.

For LDA model, we want to see whether various research topics can be identified by the model. For the purpose, we used SU = "computer science" and TS = "operating system"; SU = "computer science" and TS = "information security"; SU = "computer science" and TS = "artificial intelligen*"; SU = "computer science" and TS = "computer graphic*"; SU="computer science" and TS = "software engineer*" as search strategies to search and download bibliographic data from Science Citation Index Expanded (SCI-EXPANDED) database. The data relates to 28421 research papers in five research topics of computer science, i.e., operating system (OS), information security (IS), artificial intelligence (AI), computer graphics (CG) and software engineering (SE). Then we extracted keywords from the data and input them to the LDA algorithm. By setting the topic number as 5, we got a paper-topic matrix of 28421 rows and 5 columns.

Fig. 5 demonstrates the topic distribution of a few of papers (21 papers) as an example. It can be seen that all the papers to some extent relates to all the topics. To verify the LDA model's ability of identifying topics, we assigned the LDA topic with largest probability as a paper's research topic (e.g., the topic of *P*1 is *T*1) and recorded all the correspondences between the papers (in various research topics) and the LDA topics. The results illustrated in Table 1 show the correspondences of AI to *T*4, CG to *T*0, IS to *T*2, OS to *T*1 and SE to *T*3. Table 2 shows the 5 most representative keywords of each topic. From the table we can see that the topics can be easily interpreted by a field expert, and the interpretation in the table can perfectly support the correspondences in Table 1.

	T1	T2	T3	T4	T5
P1	0.45883832	0.164114959	0.163623793	0.075640947	0.13778198
P2	0.210309339	0.111883395	0.105178201	0.110418023	0.462211043
P3	0.104355318	0.338316175	0.138906738	0.128575894	0.289845874
P4	0.119028625	0.157018074	0.123554964	0.137053832	0.463344506
P5	0.20472695	0.176076219	0.122768193	0.12775185	0.368676788
P6	0.145228182	0.140159388	0.151644473	0.407280753	0.155687204
P7	0.144154499	0.155220457	0.165255431	0.20741911	0.327950503
P8	0.14415475	0.155218519	0.165255089	0.207421455	0.327950188
P9	0.166451482	0.131208692	0.179222426	0.304441634	0.218675766
P10	0.237715337	0.136623345	0.170003951	0.229389012	0.226268355
P11	0.153039434	0.18518332	0.172238949	0.254348291	0.235190006
P12	0.174117715	0.153762479	0.176797148	0.270478613	0.224844045
P13	0.265895703	0.176708621	0.236357873	0.173821799	0.147216004
P14	0.150518304	0.151840357	0.200534308	0.26336887	0.23373816
P15	0.159640511	0.135728359	0.275412381	0.20493017	0.224288578
P16	0.130415026	0.162874373	0.198218203	0.289108154	0.219384243
P17	0.130413831	0.162875685	0.198217758	0.289109677	0.21938305
P18	0.130414759	0.162873764	0.198219831	0.289106614	0.219385032
P19	0.122749078	0.166316765	0.154509761	0.09672068	0.459703715
P20	0.158859825	0.098189067	0.254627364	0.140891816	0.347431928
P21	0.102068778	0.099542631	0.454964661	0.100606581	0.242817348

Figure 5 An example of paper-topic matrix

Table 1 Correspondences between the papers and LDA topics

	***T*0**	***T*1**	***T*2**	***T*3**	***T*4**
AI	821	457	1018	838	4383
CG	2490	69	66	140	113
IS	409	86	2226	581	211
OS	1103	1962	1897	148	125
SE	187	2804	551	5047	689

Table 2 Most representative keywords of each topic

***T*0**	***T*4**	***T*3**	***T*1**	***T*2**
Graphics	Artificial	Software	System	Security
Image(s)	Intelligence	Engineering	Operating	System(s)
Virtual	Learning	System(s)	Performance	Network
Surface	Knowledge	Project	Application(s)	Risk
Simulation	Neural	Tool(s)	Hardware	Attack(s)

For the effectiveness of competitiveness evaluation, we used AI ranking as example and compared the ranking resulted from our method with two other rankings. AI was chosen since there are numbers of public available rankings released by various third parties that can be compared. We reviewed three rankings, i.e., US News ranking 2016, CS Ranking 2016, and WUZHEN Ranking 2016[†]. We found that the three rankings were similar with each other. For example, the Top 3 institutions in all the 3 rankings are Stanford University, Carnegie Mellon University and MIT. The Top 10 institutions in the 3 ranking also have many overlapping: in total only 16 institutions make up the 3 Top 10 lists. That is, in the 30 institutions of the 3 Top 10 lists, more than 40% of the institutions are overlapped. Consequently, we chose the most widely acknowledged ranking, i.e., US News ranking 2016, as our first reference for comparison. For the other reference, we want to compare the topic based method with the traditional method. That is, after topic recognition, we removed the concept of multi-disciplines from the paper-topic matrix Θ so that each paper only contributes to the topic with largest probability. More formally, $\forall \theta_{ij} \in \Theta, \theta_{ij} = 1$, if $\forall j \neq k, \theta_{ij} > \theta_{ik}$; $\theta_{ij} = 0$, else. For example, the distribution of *P*1 on *T*1 to *T*5 in Fig. 5 will be 1, 0, 0, 0, 0. The following steps were all the same with the proposed method and the results were acquired for comparison. During the case study, citation counts and impact factors of papers were used for paper-impact vector calculation, and weights were set as 0.5 and 0.5. The results are presented in Table 3.

From Table 3 we can conclude that the topic based method is more effective than the traditional method for at least the following 2 aspects. First, 50% of the Top 10 institutions resulted by the proposed topic based method are the same as US News Ranking, much higher than the 20% of those by the traditional method. Second, for those *toppest* institutions, 2 of the Top 3 institutions resulted by the proposed topic based method are also in the Top 3 in the US News Ranking, and the Carnegie Mellon University is also ranked in Top 10. While for traditional method, none of the Top 3 institutions in the US News Ranking appears in the Top 3 list, and only Stanford University appears in the Top 10 list.

[†] https://www.usnews.com/best-graduate-schools/top-science-schools/artificial-intelligence-rankings
http://www.askci.com/news/chanye/20160816/11231254039.shtml
tech.163.com/photoview/6PGI0009/13525.html

Table 3 Comparison of the 3 Rankings

RANKING	USNEWS	TOPIC BASED	TRADITIONAL
1	Stanford University	MIT	University of Texas at Austin
2	Carnegie Mellon University	Stanford University	Nanyang Technological University
3	MIT	Nanyang Technological University	Sydney University of Science and Technology
4	UC Berkeley	Microsoft	Washington University
5	University of Washington	Chinese Academy of Sciences	National Taiwan University of Science and Technology
6	Georgia Institute of Technology	Texas State University, Austin	Stanford University
7	University of Illinois at Urbana - Champaign	Carnegie Mellon University	Valencia University of Technology
8	Texas State University, Austin	University of London	Carlos III University of Madrid
9	Cornell University	Castilla-La Mancha University	Shanghai Jiaotong University
10	University of California at Los Angeles	University of Illinois at Urbana - Champaign	Indian Institute of Technology

Finally, by setting $\tau = 0.15, p = 10\%$, then MIT, Stanford University, Nanyang Technological University, Microsoft, Chinese Academy of Sciences, Texas State University-Austin, and Carnegie Mellon University are leading the research on AI. While Beijing University of Posts and Telecommunications, Northeastern University, Wuhan University, Graz Technical University, Paderborn University, Malaysia University of Technology and many other institutions still need to improve their research on AI.

Conclusion

This paper presents a method of evaluating the competitiveness of research institutions based on research topic distribution. The method uses the LDA topic model to obtain paper-topic distribution matrix to objectively assign the academic impact of papers (such as times of citation) to research topics. Then the method calculates the competitiveness of each research institution on each research topic with the help of institution-paper matrix. Finally, the competitiveness and the research strength and/or weakness of the institutions are defined and characterized. Case study shows that the method can lead to an objective and effective evaluation of the research competitiveness of given research institutions on given research field.

It is to be noted that the proposed method formally defines the process competitiveness evaluation, i.e., topic analysis, impact allocation and competitiveness measurement. Many adaptations can be easily applied in practice. For example, one can integrate the process of impact allocation with expert preference by multiplying Θ^T with a weight vector. In the meanwhile, more indicators such as centrality of a paper in the citation network can also be considered for paper impact characterization. Besides, since LDA model can handle topics on multi-granularity, by considering different research bodies (e.g., countries, institutions or researchers), the method can be easily used to evaluate the competitiveness of research bodies and topics on various levels of details. Our future work will consider the mentioned adaptations for a more effective competitiveness evaluation. Other work like mining patterns among strengths and/or weaknesses may also be considered so that the relationships among topics can be further understood.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant (No. 7160325); the Young Talent-Field Frontier Project of Wuhan Documentation and Information Center, Chinese Academy of Sciences.

References

Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

Braam R, Moed H F (1991). Mapping of science by combined co-citation and word analysis: structural aspects. Journal of the American Society for Information Science, 42(4): 233-251.

Chen, Chaomei. CiteSpace II (2006). Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for information Science and Technology, 57.3: 359-377.

Chen, Chaomei, Zhigang Hu, Shengbo Liu, and Hung Tseng (2012). Emerging trends in regenerative medicine: A scientometric analysis in CiteSpace. Expert opinion on biological therapy, 12 (5): 593-608.

Chen Shiji, Shi Liwen, Zuo Wenge (2013). Theoretical and Empirical Study on Measure Method of Academic Influence Indicator Based on ESI. Library and Information Service, 57(2): 97-102.

Cova, T ânia FGG, Alberto ACC Pais, and Sebasti ão J. Formosinho.. Iberian Universities: a Characterization from ESI Rankings. Scientometrics，2013，94(3): 1239-1251.

Dong Zheng'e, Chen Huilan (2014). Investigation into Library Service Model of University Discipline Evaluation on the Basis of ESI and InCites Databases [J]. Library Journal, 33(11): 23-28.

Gei Fei, Tan Zongying (2013). Emerging Trend Detection Methods of the Subject Discipline Area Themes. Information Studies:Theory & Application, 36(9):78-82.

Li Maomao，Zhang Ziqian，Chen Shiji，Zuo Wenge (2012). An ESI based Analysis of the Competitiveness in Plant & Animal Sciences of China Agricultural University [J]. Science and Technology Management Research, (8): 128-132.

Liu, Jianhui, and Mei Ye (2015). "Research on the Subject Development Forecast Based on ESl and lnCites--Taking China University of Geosciences as an Example." Sci-Tech Information Development & Economy, 6: 056.

Mkhnacheva Y (2011). Research Performance of RAS Institutions and Russian Universities: A Comparative Bibliometric Analysis. Herald of the Russian Academy of Sciences, 81(6): 569-574.

Morris S A, Yen G, Wu Z (2003). Time line visualization of research fronts. Journal of American Society for Information Science, 54(5): 413-422.

Small H, Upham P (2009). Citation structure of an emerging research area on the verge of application. Scientometrics, 79(2): 365-375 .

Shibata N, Kaiikawa Y, Takdea Y (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. Technovation, 282(11):758-775.

Small, Henry, Kevin W. Boyack, and Richard Klavans (2014). Identifying emerging topics in science and technology. Research Policy, 43(8): 1450-1467.

Zhang Faliang, Tan Zongying, Wang Yanping (2014). Measurement of the Research Topics of Research Institutions: A Case Study on Information Science of China. Library and Information Service, 58(8):85-90.