Expertise-based institutional collaboration recommendation in different thematic areas

Hiran H. Lathabai, Abhirup Nandy and Vivek Kumar Singh

Department of Computer Science, Banaras Hindu University, Varanasi, Uttar Pradesh-2210005, India.

Abstract

Institutions are known to have varying research strengths in different thematic areas. While some thematic areas are within the core competence of an institution, there may be other areas in which the institution is considered relatively week. This work proposes an expertise-based recommendation framework that can determine the stronger and weaker thematic areas of an institution based on their expertise and toss recommendations. The framework uses bibliometric and text data and applies methods from Network Science and Text Analytics. The recommendations provided can be useful for various purposes ranging from suggestions for institutional collaborations for improving an institution's research performance in a weaker thematic area (by pairing with an institution stronger in the corresponding thematic area) to research place recommendations to prospective researchers. This unique capability of the framework is demonstrated using 196 research institutions in India. Results are compared with available evidence from different international rankings and the ability of the framework to provide novel recommendations is established.

Keywords

Expertise determination, Institutional recommendation, Research place recommendation, *x*-index, core competency, thematic strength.

1. Introduction

Three organizations of science, viz. the social, institutional and intellectual organizations are interwoven, and strongly influence each other for the cause of progress of science. Innovative research usually is initiated through one or more institutions and taken up by competent researchers from other institutions. In this way, the institutional organization of research involves in the progress of science through the influence of and by influencing other two organizations of science. A global shift from 'trust-based funding' to 'performance-based funding' [16] forced funding agencies to adopt sharp performance assessment mechanisms and institutions to adopt schemes to keep their performance steadily upright. However, due to many factors, with respect to a single discipline/field, an institution always has some strong areas of research that can be regarded as their core competency areas. For a predominantly large number of other sets of thematic areas, that institution might have a relatively weak research performance. In such cases, the respective institution may be interested to collaborate with other institutions which would be core-competent in such a thematic area(s). A mechanism to determine both the aspects is vital for research institutions to sustain their prestige and rise towards excellence. Recommendation or recommender systems can bridge this gap to a great extent. The question, however, is whether the existing recommendation systems sufficiently addressed this problem and delivered promising results.

Collaboration in science is one of the well-explored topics since its inception as a response to professionalization of science [4], especially in the form of author collaboration networks or co-authorship networks. Some of the earlier attempts for measurement/assessment of collaboration through co-authorship revealed (i) the need to survey and follow up the issues of collaboration, (ii) how various aspects of collaboration can be analyzed through refined use of co-authorship

BIR 2021: 11th International Workshop on Bibliometric-enhanced Information Retrieval at ECIR~2021, April 1, 2021, online
EMAIL: hiranhl007@gmail.com (A. 1); abhirupnandy.online@gmail.com (A. 2); vivek@bhu.ac.in (A. 3)
ORCID: 0000-0002-5633-9842 (A. 1); 0000-0001-8618-0847 (A. 2); 0000-0002-7348-6545 (A. 3)

bibliometrics [13], and (iii) the limitations of co-authorship-based studies due to research culture and practices among individuals or organizations in different disciplines [9]. These studies however, acknowledged the effectiveness of co-authorship networks in investigation of patterns of scientific collaboration and its dynamics. Major works on co-authorship networks for recommendation include link prediction approach that uses (i) node features or attributes like common neighbors [14] or fractionally counted common neighbors [1], Resource Allocation index or RA-index [17] and (ii) structural aspects of network(s) like structural similarity indies [12]. Co-authorship link prediction framework using multiple relations in scientific literature modelled as multiplex networks by Pujari [18] and link semantic framework by [5] using semantic features are found to be effective. However, co-authorship recommendation systems do not address the above-mentioned problems and are not suitable for institutional collaboration recommendations.

There are some studies that (i) explored the evolution patterns in co-institution networks based on derived co-occurrence matrix from publication-institution matrix [19], (ii) mapped the institutional collaboration networks for identification of most collaborative institutions within research fields [2, 10, 20]. Some other exploratory works have dealt with the effects of inter-firm collaborations which include university-industry collaborations [6,7,15,22]. To the best of our knowledge, institutional collaboration recommendation in academia, based on thematic research areas, is almost an unexplored field. To bridge this gap, we introduce a preliminary form of a recommendation system (that may be developed into a full-fledged one) based on institution's expertise/core competency.

2. Recommendation framework based on thematic strength and core competency

As discussed earlier, our recommendation framework has two parts- (i) expertise determination for identifying research strength of an institution, and (ii) recommendation retrieval for weak performing research areas of an institution. Schematic diagrams of the framework showing both the sections can be found in **Figure 1**.

2.1. Expertise determination

The proposed framework based on expertise of institutions in thematic areas makes use of an index named *x*-index, which is designed by adopting the notion of *h*-index [8]. The definition of *x*-index is given as follows:

x-index: An institution is supposed to have an *x*-index value of *x* if it has published papers in at least *x* thematic areas and received at least *x* citations in those areas. These *x* areas that form the *x*-core can be treated as the core competency areas of the institution. High value of *x* indicates that institution has got expertise/competency in more diverse thematic areas.

As a first step of our framework, publication data of institutions in a field is collected. In scientific publications, most of the journals prompt authors to specify keywords as an attempt to signify the specific contribution of that article and also as an attempt to identify the thematic areas within which an article falls. In this work, such author-provided keywords (field with tag 'DE' in Web of Science or WoS file) are used to designate thematic areas following the investigation by [21] that implied the effectiveness of authorfor investigation of knowledge structure within scientific fields.

We use a network-based approach to map the publication-keyword relationship, in the form of an affiliation-network [3]. For the same, a Work-Keyword (Author) or W-K(A) network is built. The publications and the author provided keywords are two sets of nodes in the network or graph, with a directed link *l* between the nodes *a* & *b* if a publication (or work),denoted as *a*, consists of a keyword, denoted as *b*. For the next segment, we convert the W-K(A) network to a weighted W-K*(A) network by following two steps –

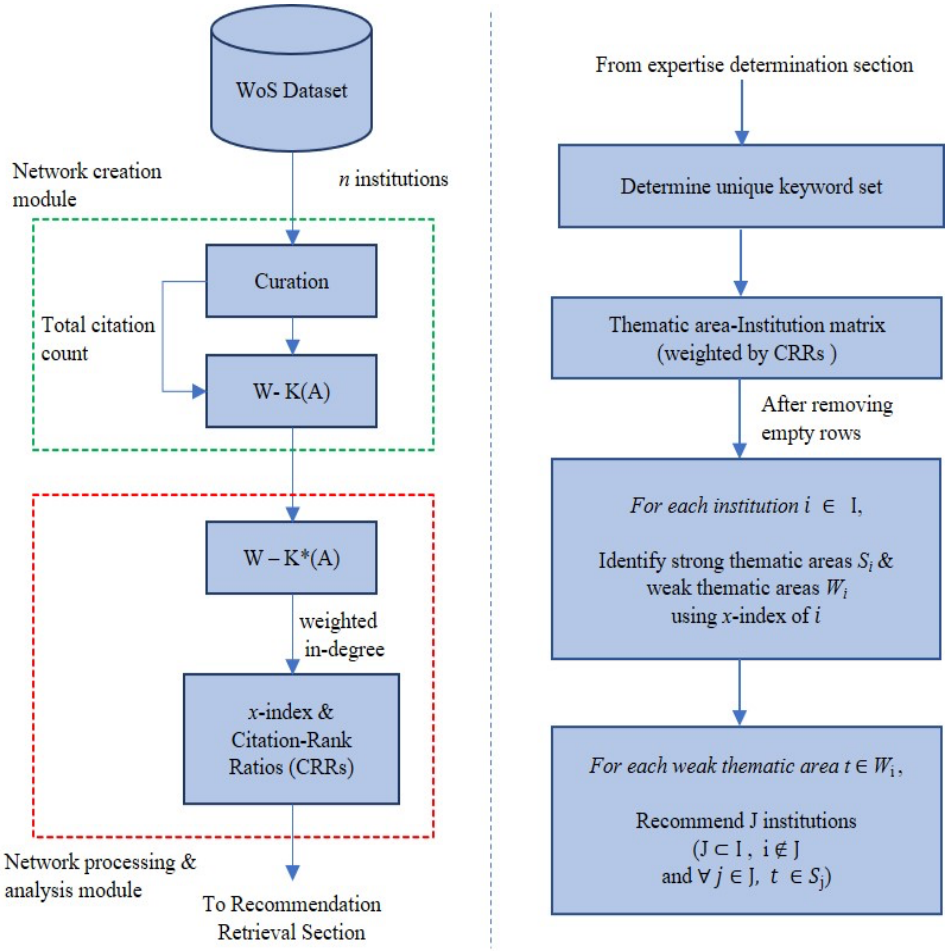

Figure 1: Schema of the recommendation system

1. Data curation step, by eliminating bad or missing valued data (the fields checked are citation count and author keywords).

2. Link (edge or arc) weightage calculation step, by calculating the weight *f* of a link from publication-work *w* to keyword *k*, where *f* = number of citations of *k* by the virtue of *w* (can be found in field with tag 'Z9' in WoS file).

The conversion of the W-K(A) to W-K*(A) network achieved by this module is equivalent to the 'injection process' which is introduced as part of the 'injection methodology' by [11].

The computation of *x*-index is a major task for this study. For the keyword nodes, we calculate the weighted-indegree value, which corresponds to the total number of citations received for each keyword. The keywords are then sorted and ranked according to the citation values. Now the *x*-index is computed in a *h*-index like fashion, by computing the Citation-Rank-Ratio (CRR) and identifying the point where the CRR value crosses unity. Formally, *x* is the first occurrence of one of the following cases:

$$x = \begin{cases} r, & if\ \text{CRR} = \dfrac{citation\ at\ position\ r}{r} = 1 \\ r-1, & if\ \text{CRR} = \dfrac{citation\ at\ position\ r}{r} < 1 \end{cases}$$

So, the *x*-core of an institution is demarcated by the CRR values in the following way:

1. A keyword *k(a)* belongs to the *x*-core if CRR ≥ 1
2. Otherwise, *k(a)* belongs to the *x*-tail

The *x*-index value of an organization represents the core competency of an institution in a field represented by keywords and *CRR* values that demarcate the core thematic areas and the rest can be useful for retrieving the recommendations for institutions. Both of this information is processed by the next section of our recommendation framework, namely the 'recommendation retrieval' section.

2.2 Recommendation retrieval

Let our set of institutions be *I*, where $\forall\ i \in I$, the previous section provides a set of keywords k_i. So, we have a unique set of keywords *K*, which can be identified by $K = k_1 \cup k_2 \cup k_3 \ldots \cup k_N$, where $N = |I|$. The second module then creates a Keyword – Institution (K-I) matrix of size $M \times N$, where $M = |K|$, with the keywords and institutions creating the rows and columns respectively. The values of the matrix are filled with the corresponding CRR values as shown -

$$\mathrm{K-I} = \begin{pmatrix} \mathrm{CRR}_{11} & \cdots & \mathrm{CRR}_{1N} \\ \vdots & \ddots & \vdots \\ \mathrm{CRR}_{M1} & \cdots & \mathrm{CRR}_{MN} \end{pmatrix}$$

The rows with keywords uncited for every institution are eliminated, thus resizing our K-I matrix to $M' \times N$.

Now, for each institute $i \in I$, we check which of the keywords $k \in M'$ lie in the *x*-tail region of the institute. All those keywords are marked to be in the weak thematic area W_i of the institution, whereas the ones in *x*-head falls under the strong thematic area S_i of the institution.

Lastly for out final segment of the recommendation system, we select an institution *i*, and for a weak thematic area W_i, we recommend a list of institutions J, where for each keyword $k \in W_i$, $k \in S_j$ such that $j \in J$ and $i \neq j$. Thus, for an institution *i,* for the thematic areas in which *i* is relatively weak, institutions with relatively high expertise are selected to be suggested. Now, theorder of recommendations can be based on the total number of citations received by keyword $k \in S_j$ for an institution $j \in J$.

3. Data collection

Data collection is done from Web of Science (WoS) database, which is one of the largest online databases that indexes scholarly documents from reputed sources and thereby regarded as a standard database for bibliometric research. The dataset consists of institution-wise scholarly publications, within the time-period of 2010 to 2019. Computer Science was chosen as the discipline/subject for which data collection was carried out and the dataset included every type of documents in the database. Only those institutions are selected, which had at least of 25 publications within the period. A total of 196 Indian institutions (excluding institution systems like CSIR, IIT systems etc.) satisfied the given criteria. The metadata fields considered for the data collection are Author Keywords (DE) and Total Times Cited Count (Z9) of each of the distinct publications.

4. Results and discussion

From the dataset of all 196 institutions, *x*-indices of every institution with CRR values are computed using the first section of the framework. Top 10 institutions with high *x*-indices and their *x*-index values are: Thapar Institute of Engineering Technology (115), IIT Kharagpur (115), ISI Kolkata (103), IIT Delhi (97), IIT Roorkee (95), Vellore Institute of Technology (84), IIT Kanpur (77), IISc Bangalore (72), NIT Rourkela (68) and Anna University (65). Therefore, both Thapar Institute of Engineering Technology (TIET) and IIT Kharagpur gathered 115 citations or above in at least 115 areas. These 115 areas are supposed to be the core competencies of these institutions. After computing the CRR values and x-indices, we proceed to the next section.

First two modules of second section produced the updated K-I matrix. There were 46,859 unique keywords (which were at least cited once). For demonstrating the framework, we selected TIET, the institution that tops in the list of institutions with high expertise index. From the CRR matrix, we have identified all the thematic areas in which TIET is relatively weak compared to other thematic areas with high citation counts (115 out of 46,859 are strong and the rest are either relatively weak or absolutely weak areas), and arbitrarily a selected one thematic area from the weak ones for demonstration. This area is 'Data mining' and its CRR value (rounded to three decimal places) is 0.099.

Now, the institutions that are strong in these thematic areas can be identified using the CRR values of institutions (with respect to these thematic areas). Institutions with CRR values ≥ 1 are strong in these areas and they can be recommended for TIET. For Data mining, total number of institutions that can be recommended for TIET is 15. The priority/order of recommendations is decided using thematic strengths of the institutions in these areas (reflected by the number of citations). Top 10 institutions that can be recommended to TIET in areas Data mining is shown in **Table 1.**

Table 1.

Top 10 institutions that can be recommended to TIET for the area 'Data mining'

Rank	Institutions recommended to TIET for area 'Data mining'	Citations
1	ISI KOLKATA	205
2	IIT KHARAGPUR	140
3	GAUTAM BUDDHA UNIVERSITY	138
4	MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY JAIPUR	138
5	GURU GHASIDAS VISHWAVIDYALAYA	123
6	IIT INDORE	119
7	JAWAHARLAL NEHRU UNIVERSITY NEW DELHI	119
8	IIM AHMEDABAD	114
9	NIT SILCHAR	91
10	IIEST SHIBPUR	86

The framework for recommendation system developed in this work is quite different from the existing kinds of recommendation systems in academia. Therefore, a comparative validation of its performance based on existing techniques/measures is not possible. We have also checked the possibility of exploring whether there is an essential difference/similarity between priority/order in which our recommendation is tossed and the existing well known ranking schemes. For comparison, firstly the common institutions between Indian institutions listed in popular ranking scheme and institutions recommended in three thematic areas (separately) have to be found out. In most of the popular international overall ranking schemes of institutions QS, THE, ARWU and CWTS Leiden, Indian institutions are underrepresented of these due to many factors. Therefore, common institutions will be even less in number and with small data size, significantly indicative results cannot be obtained. Further, it may be noted that these international rankings provide ranks in an overall subject only (Computer Science in this case) and not in specific thematic areas. Among the above four ranking schemes, relatively better representation of Indian institutions (6 common institutions for Data mining) is found in CWTS Leiden. Therefore, we compared the relative CWTS Leiden ranks and relative priority/order of our recommendations of common institutions using Spearman's rank correlation. Spearman's ρ for the area Data mining is 0.229. The same procedure is followed for comparison of our recommendation order with National ranking framework (NIRF) of India. Number of common institutions for NIRF and priority/order of recommendations for the area Data mining is 8. Therefore, for NIRF, Spearman's ρ for the area Data mining is 0.155. As rank correlation strengths are found to be weak (< 0.9), our recommendation results are found to be sufficiently different from the existing ranking schemes. This

indicates the ability of our framework to toss novel recommendations by ruling out the obviousness or possibility of easy accessibility of our recommendations from existing rankings.

5. Conclusion

Institutional performance improvement is vital for the concerned institution as well as the country in which they belong to. Consequently, the fields of research in which that institution is engaged upon can be benefited. One major step for improving the performance of institutions is the assessment of institutional performance to know the present position of institution. Second major requirement is the actual efforts to improve the position of institution by increasing productivity/impact of scholarly output. As collaborative research is known to be effective for improving the productivity/impact of institutions, identification of suitable partner institutions to collaborate can be regarded as a key strategy for performance improvement. For assessment of institutional performance and identification of the overall position and the position of institutions in broad subject categories there are so many ranking frameworks. However, frameworks that can identify the research strength of institutions at thematic area/ fine subfield level and compute expertise of the institution based on these are almost rare. When it comes to identification of suitable partner for collaboration, recommendation systems can be extremely useful. However, existing recommendation systems in academia are mostly based on co-authorship of individuals. Though there are a few studies on co-institutional relationship patterns using networks, institutional recommendation system (that too based on research area strengths) is almost unexplored. This gap is attempted to be bridged by the development of an institutional recommendation system that can identify the strength of institutions in different thematic areas within a discipline/subject and thereby the areas in which it is strong and weak and capable of recommending institutions that are strong in the areas in which a particular institution is weak. The identification of strong and weak areas is achieved by the help of expertise index or *x*-index, which is computed in *h*-index like fashion but takes into consideration, the strengths of institutions (reflected by citations) in each thematic area. Citations in each thematic area are computed using network approach by the first section of the recommendation system. Second section retrieves the suitable recommendations for an institution in a thematic area. The framework, which is the first of its kind, is demonstrated using data for one Indian institution (among 196) with the highest expertise. Three thematic areas in which that institution is relatively weak are selected and our framework is found to be capable of retrieving institutions that have high research strength in these areas. Upon comparison with a major international ranking scheme and a national ranking scheme using Spearman's rank correlation, the novelty of the recommendations and the ability of our framework to toss novel recommendations is established.

A major limitation of the framework lies in its dependency on author keywords for determination of thematic areas. So, the accuracy of our framework is very much dependent on how well the author keywords represent thematic areas in a field. This can be improved by using 'Natural Language Processing' (NLP) in the pre-processing phase of the framework. For instance, with NLP, singular and plural versions of keywords can have a representative term, and too generic terms like 'Model', 'Method', etc., can be eliminated. This is intended to be pursued as a future endeavor to improve this framework. Another possible exploration is the usage of advanced network analysis techniques to gather more insights from the framework.

6. Acknowledgments

The authors would like to acknowledge the support provided by the DST-NSTMIS funded project-'*Design of a Computational Framework for Discipline-wise and Thematic Mapping of Research Performanceof Indian Higher Education Institutions (HEIs)*', bearing Grant No. DST/NSTMIS/05/04/2019-20, for this work.

7. References

[1] L. A. Adamic, E. Adar, Friends and neighbors on the web, Social networks 25.3 (2003): 211-230. https://doi.org/10.1016/S0378-8733(03)00009-1
[2] I. Ashraf, S. Hur, Y. Park, An Investigative Analysis on Finding Patterns in Co-Author and Co-Institution Networks for LIDAR Research, INTERNATIONAL ARAB JOURNAL OF INFORMATION TECHNOLOGY 17.6 (2020): 875-884. https://doi.org/10.34028/iajit/17/6/6
[3] V. Batagelj, Social Network Analysis, Large-Scale, in A. Robert Meyers (Ed.), Computational complexity: Theory, techniques, and applications,New York: Springer, 2009, pp. 2878–2897.
[4] D. Beaver, R. Rosen, Studies in scientific collaboration: Part I. The professional origins of scientific co-authorship, Scientometrics 1.1 (1978): 65-84. https://doi.org/10.1007/BF02016840
[5] M. A. Brandão, M. M. Moro, G. R. Lopes, J. P. Oliveira, Using link semantics to recommend collaborations in academic social networks, in Proceedings of the 22nd International Conference on World Wide Web, 2013, pp. 833-840.
[6] K. Chen, Y. Zhang, G. Zhu, R. Mu, R. Do research institutes benefit from their network positions in research collaboration networks with industries or/and universities? Technovation94 (2020): 102002. https://doi.org/10.1016/j.technovation.2017.10.005
[7] J. Guan, Q. Zhao, The impact of university–industry collaboration networks on innovation in nanobiopharmaceuticals, Technological Forecasting and Social Change 80.7 (2013): 1271-1286. https://doi.org/10.1016/j.techfore.2012.11.013
[8] J. E. Hirsch, An index to quantify an individual's scientific research output, Proceedings of the National academy of Sciences, 102 (46) (2005), pp.16569-16572.
[9] J. S. Katz, B. R. Martin, What is research collaboration? Research policy 26.1 (1997):1-18. https://doi.org/10.1016/S0048-7333(96)00917-1
[10] M. A. Koseoglu, Mapping the institutional collaboration network of strategic management research: 1980–2014,Scientometrics 109.1 (2016): 203-226.https://doi.org/10.1007/s11192-016-1894-5
[11] H. H. Lathabai, T. Prabhakaran, M. Changat, Contextual productivity assessment of authors and journals: a network scientometric approach. Scientometrics, 110.2 (2017): 711-737. https://doi.org/10.1007/s11192-016-2202-0
[12] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, Journal of the American society for information science and technology 58.7 (2007): 1019-1031. https://doi.org/10.1145/956863.956972
[13] G. Melin, O. Persson, Studying research collaboration using co-authorships, Scientometrics36.3 (1996): 363-377.https://doi.org/10.1007/BF02129600
[14] M. E. Newman, Coauthorship networks and patterns of scientific collaboration. Proceedings of the national academy of sciences, 101, 2004, pp.5200-5205. https://doi.org/10.1073/pnas.0307545100
[15] M. A. Schilling, C. C. Phelps, Interfirm collaboration networks: The impact of large-scale network structure on firm innovation, Management science 53.7 (2007): 1113-1126.
[16] S. Sörlin, Funding diversity: performance-based funding regimes as drivers of differentiation in higher education systems. Higher Education Policy 20.4 (2007):.413-440. https://doi.org/10.1057/palgrave.hep.8300165
[17] Q. Ou,Y. D. Jin, T. Zhou, B. H. Wang, B. Q. Yin, Power-law strength-degree correlation from resource-allocation dynamics on weighted networks, Physical Review E 75.2. (2007):021102. https://doi.org/10.1103/PhysRevE.75.021102
[18] M. Pujari, R. Kanawati, Link prediction in multiplex networks. Networks &Heterogeneous Media 10.1 (2015): p.17. doi: 10.3934/nhm.2015.10.17
[19] C.S. Wagner,&L. Leydesdorff, Mapping the network of global science: comparing international co-authorships from 1990 to 2000. International journal of Technology and Globalisation, 1(2) (2017): 185-208.https://doi.org/10.1007/s11192-016-1894-5
[20] Q. Ye, H. Song, T. Li, Cross-institutional collaboration networks in tourism and hospitality research. Tourism Management Perspectives, 2 (2012): 55-64. https://doi.org/10.1016/j.tmp.2012.03.002
[21] J. Zhang, Q. Yu, F. Zheng, C. Long, Z. Lu, Z. Duan, Comparing keywords plus of WOS and author keywords: A case study of patient adherence research, Journal of the Association for Information Science and Technology 67.4 (2016): 967-972. https://doi.org/10.1002/asi.23437

[22] Rupika, A. Uddin, V.K. Singh, Measuring the University-Industry-Government Collaboration in Indian Research Output. Current Science, 110(10) (2016): 1904-1909.