https://doi.org/10.1007/
s11192-021-04188-3

x-index: Identifying core competency and thematic research strengths of institutions using an NLP and network based ranking framework

Hiran H. Lathabai[1] · **Abhirup Nandy**[1] · **Vivek Kumar Singh**[1]

Received: 25 April 2021 / Accepted: 11 October 2021 / Published online: 8 November 2021

Abstract
The currently prevailing international ranking systems for institutions are limited in their assessment as they only provide assessments either at an overall level or at very broad subject levels such as Science, Engineering, Medicine, etc. While these rankings have their own usage, they cannot be used to identify best institutions in a specific subject (say Computer Science) by taking into account their performance in different thematic areas of research of the given subject (say Artificial Intelligence or Machine Learning or Computer Vision etc. for the subject Computer Science). This paper tries to bridge this gap by proposing a framework that uses the NLP and Network approach for identifying the *core competency* of institutions and their *thematic research strengths*. The *core competency* can be viewed as a measure of breadth of research capability of an institution in a given subject, whereas *thematic research strength* can be viewed as depth of research of the institution in a specific theme of a subject. The working of the framework is demonstrated in the area of Computer Science for 195 Indian institutions. The framework can be useful for institutions and the scientometrics research community as a system providing a detailed assessment of the core competency and the research strengths of institutions in different thematic areas. The framework and outcomes can also be useful for funding agencies in devising programs for 'performance-based funding' in 'thrust areas' or 'national priority areas'.

Keywords Core competency · Expertise indices · Research strength · Thematic strength · *x*-index

Introduction

Scientific progress of a nation can be viewed as a self-organization of its intellectual, social and institutional constructs through complex dynamics. Institutional organization of science has now become the norm for doing science. Institutions obtain their resources for research from various agencies, though majority part in most of the countries still comes

from public funding. Due to limits of public funding, governments and funding agencies are getting compelled to adopt a differential funding approach through the use of some kind of performance-based systems. In fact, a major shift from 'trust-based funding' to 'performance-based funding' was happening in the last few decades among various countries. Major reason for the adoption of performance-based funding was the hope for promotion of 'vertical differentiation' and 'functional specialization' between the institutions and the simultaneous assurance of horizontal diversity and pluralism within the system (Sörlin, 2007). The formation of Research Excellence Framework (REF) in UK (De Boer et al., 2015), Australian government's decision to allocate AUD 80 million to universities based on performance-based funding (Maslen, 2019), presence of 'performance-based funding' as a key component in the Norwegian model (Sivertsen, 2016) adopted at national level by Belgium (Flanders), Denmark, Finland, Norway and Portugal, etc., are some of the major performance-based funding initiatives. The shift towards performance-based funding increased the relevance of performance assessment at almost all the levels like individual researcher, research group and institutional level.

Several national and international funding agencies often rely on different global ranking systems like the QS World University Rankings, Times Higher Education World University Rankings (THE), Academic Ranking of World Universities (ARWU), Centre for Science and Technology Studies (CWTS) Leiden ranking, etc., for institution performance assessment for performance-based funding. Majority of the above-mentioned global ranking systems are found to use different kinds of survey data, along with information such as (i) number of publications in some highly reputed journals, (ii) achievements of the faculty and alumni like Nobel Prizes and Fields Medals (Shanghai Ranking's Academic Ranking of World Universities, 2021), and/or (iii) other information such as grants. These methodologies, although being very popular, suffer widely from bias and major criticism. A critical approach (Billaut, Bouyssou, &Vincke, 2010) shows that the criteria used in ARWU are not relevant and that the aggregation methodology suffers from major problems. Another criticism (Jeremic et al., 2011) shows that changing the relative weights placed upon each of the six factors of ARWU significantly alters the ranking. Similarly, THE rankings, being "arguably the most influential" (Beck & Morrow, 2010), has an anchoring effect on the ratings (Bowman & Bastedo, 2011), which is extensively exploited by institutions to represent themselves. QS Rankings on the other hand, has commercialized the ranking system by inclusion of star ratings, along with including excess emphasis on peer review (Anowar et al., 2015). These international rankings further have the problem of inclusion, more so for institutions in developing countries. This is why several countries have developed national ranking frameworks. When it comes to national level ranking systems, there are several interesting examples in UK, Australia, Italy, Norway, Spain etc. The recently introduced National Institutional Ranking Framework (NIRF) in India has a similar purpose. However, none of these ranking systems provide for assessment of institutional research performance at the level of thematic areas or to determine its core competency areas.

In this era, as the world witnesses the emergence of more and more fields including the interdisciplinary ones at a quicker pace, countries are looking for national strategies for developing institutions of excellence in thrust areas. For instance, in India, cyber security, multiscale modeling, quantum theory and applications, etc., are some of the thrust areas in engineering sciences identified by a working group on thrust areas in 2006. A major concern that arises in this context is—"Are the existing ranking methods/systems suitable and sufficient for evaluation of institutions for performance in thrust areas?" As we see in the 'Related Work' section, there is apparently no or very limited effort in identifying research strengths of institutions in fine-grained thematic areas, and/ or determining

their core competency in research. Therefore, a framework with the capability to reflect the thematic strengths of an institution too, while assessing its performance within a field is necessary. Such a framework can be used to: (i) select institutions for funding in a specific thematic area, so as to eventually develop these as centres of excellence, and (ii) identify top performers in a given thematic area, which can help in several science policy related decisions. Motivated by this, we propose a framework that can determine the core competency of an institution in a given subject and also determine its thematic research strength in different themes of that subject. In some sense, the *core competency* can be viewed as a measure of breadth of research capability of an institution in a given subject (say Computer Science), whereas the *thematic research strength* can be viewed as depth of research of the institution in a specific theme of the subject (say Particle Swarm Optimization in Computer Science).

The main characteristic of this framework is that, unlike other global or national ranking approaches that consider the general scholarly output of institutions (that favors some institutions), performance of an institution within a field is computed based on the output of institutions in different thematic areas. Instead of subject categories in major databases that tend to be very broad representations of the possible themes within a field, thematic area determination is done based on keywords (author keywords as well as keyword plus) through a NLP-based method. This follows the finding by Zhang et al. (2016) that both author keywords and keyword plus are equally effective for investigation of knowledge structure within scientific fields. The framework we propose uses established methodologies like 'Injected direct approach' (Lathabai et al., 2017). Affiliation networks or 2-mode networks (of scientific publications and keywords) and citation networks of publications are the types of networks used in this framework. The framework is then exploited to develop indicators of expertise, which are briefly introduced next.

The major contribution of this work is the development of the expertise indices for the determination of the productive thematic core or competent thematic core and potential competent thematic core of an institutional research portfolio. The first index, which is an adoption of the notion of h-index, is named as x-index, in order to represent expertise and the second index is termed as x(g)-index as it is inspired from g-index. The proposed framework is applied to institutions to determine: (a) Core competency thematic areas (and also potential competency areas) of an institution, and (b) Thematic research strengths of an institution including the core strengths. The use of framework for identifying the core competency thematic areas and the core thematic strengths of an institution, is demonstrated using the case study of 195 institutions in India. An attempt to compare the relative rankings of Indian institutions obtained from popular rankings such as QS, CWTS, ARWU, etc., with our *x*-index based ranking is also included in this work. Such a comparison is done using Spearman's rank correlation to explore the extent of similarity or difference of our ranking framework to the existing ones.

The rest of the paper is organized as follows: First of all, a brief survey of related scientometric exercises for institutional rankings is presented in "Related work". Then, the details of the dataset used for demonstration of the framework are presented in Sect. "Data". The Sect. "Ranking framework based on thematic strength and core competency" explains in detail the proposed framework based on NLP and network-based approach for identifying core competency and determining thematic research strength of institutions. Results and analysis of results related to the demonstration of the framework are given in Sect. "Results". It is followed by a brief discussion of the proposed framework and its comparison with existing ranking approaches, in Sect. "Discussion". The Sect. "Conclusions" presents a summary of the work along with major conclusions. The

limitations of the work and some directions for possible future research are presented in Sect. "Limitations and future work".

Related work

Some efforts in scientometric literature tried to address the need for institutional performance assessment as discussed further. A study to evaluate the research performance of departments of universities using the mean *h*-index was proposed by Lazaridis (2010), though it suffers from the shortcomings of *h*-index (Hirsch, 2005) such as favoring authors with numerous collaborations (Costas & Bordons, 2007). Also, the mean *h*-index for an institution would not favor low-citation subfields of a subject which may have been emphasized. Russian Centres of Excellence and their research (Pislyakov & Shukshina, 2014), was introduced for national and international collaboration and co-authorship network mapping. Highly cited papers and co-authored papers, within the Clarivate Analytics' Essential Science Citation Indicators (ESCI) database, were taken as two factors for evaluation of such centres. A similar National level scientific research evaluation in Italy was also performed at a field level (Abramo & D'Angelo, 2014, 2020), which introduced some international comparisons for determining the research output evaluations. But such fields were broad subject categories, like Mathematics and Computer Science, or Chemistry, from the Web of Science (WoS) database. Another national level performance evaluation using a Composite index (Basu et al., 2016) was made for ranking the Central Universities of India. This index used bibliometric data to create a 'Quality-Quantity' index, to rank institutions at a national or regional level, and did not address specific thematic areas.

Some other scientometric studies focused on developing models for ranking in specific subjects. A bi-dimensional index (Torres-Salinas et al., 2011) was introduced for the fields of Chemistry and Computer Science within the Spanish Universities. This index used six different bibliometric indicators, based on WoS dataset, to calculate the scores for each institution only on the two broad categories. Another ranking of research output of universities based on the multidimensional prestige of influential fields (García et al., 2012), with flexibility to reset different thresholds and obtain different levels of ranking, was also introduced. Another approach to evaluate field specific excellence of institutions worldwide (Bornmann, Stefaner, Anegon, & Mutz, 2014) was made, but the fields were broad subject categories from Scopus database. Another study used a sciento-text approach to classify articles into 24 subfields of Computer Science and then rank the institutions (Uddin et al., 2016). Though this study used fine-grained thematic classification of research articles and tried to assess research strengths of institutions at fine-grained thematic area level, but it used the simple metric of publication count to represent research strength of institutions. Further, the thematic classes in this work were predetermined.

Most of the above discussed rankings/ studies propose either an overall rank or provide rank of institutions in different broad subject areas. These rankings, in addition to the already discussed criticisms, suffer from two main problems- (1) their design criteria often result in the exclusion or under-representation of institutions in developing countries, and (2) they do not provide a method to determine research strength of institutions in a given thematic area (say 'machine learning' in the field 'computer science') and do not consider the effect of thematic strengths in determining the performance of institutions in a field, as pointed out by López-Illescas et al. (2011). It is, therefore, important that a framework for determining core competency and thematic area-wise research strengths of institutions be

developed. Core competency and thematic research strength of an institution can together represent its research strength at a specific level of granularity. Before discussion of the proposed framework, details of data collection (that makes the first step of the framework) is discussed.

Data

The data used for this method was collected from the Web of Science (WoS) database, which is an acceptable standard in scientometric research. Data collection was done institution-wise for the country 'India', within the time-period of 2010 to 2019. Computer Science was chosen as the subject/field owing to the presence of vast and diverse research themes within it, and also due to familiarity of the authors with the subject. All document types were included for the study. Only those institutions were selected, which had at least of 25 publications in the subject within the period of 10 years. There were 195 Indian institutions (excluding institution systems like CSIR, IIT systems etc.) that satisfied the criteria. Accordingly, the publication metadata was downloaded for all these institutions. Thus, downloaded file for each institution consists of works published in the field of Computer Science by that institution during the time span 2010–2019. For instance, IIT Kharagpur has published 1205 papers in the field 'Computer Science' during 2010–2019. The metadata fields considered for the analysis and their WoS field tags (given in the fashion- field name (field tag)) were: WoS IDs of the publications (UT), Author Keywords (DE), Keyword Plus (ID) and Total Times Cited Count (Z9), for each of the distinct publications. While data in the fields with tags UT and DE are used for creation of affiliation network of publications and author keywords, data in fields with tags UT and ID are used for creation of affiliation network of publications and keyword plus. Citation scores of publications in the field Z9 is used for the conversion of unweighted affiliation networks to weighted affiliation networks. Details of the working of the framework is discussed next.

Ranking framework based on thematic strength and core competency

As the objective of this research is to develop a framework for ranking institutions based on their performance in thematic areas instead of their general scholarly performance, a method to compute the performance in thematic areas (like machine learning, Internet of things, etc.,) is the primary requirement. For that, scores that reflect the performance of an institution with respect to a thematic area have to be decided and method(s) to compute these scores are to be designed. Major concern regarding this is how to determine the thematic areas of research. Scientific literature that consists of scientific publications as basic unit can represent the body of knowledge at different levels- i.e., at the levels of (i) thematic areas/subfields, (ii) fields of research, (iii) disciplines and (iv) broad subjects. However, proper determination of the boundaries/confinements of each level in scientific literature, that is available to the analysts through various indexing systems or databases, and retrieval of relevant publications for analysis, is a very tricky affair. One has to make use of search strings that are logical combinations of certain terms known as the 'keywords' and the accuracy of the retrieval depends upon the appropriateness of the choice of keywords and their combinations. Thus, keywords can be regarded as the basic unit that imparts different levels of classification of the scientific literature. Can keywords be used

to represent thematic areas (finest level of classification of the scientific literature)? Before answering this question, one important fact needs to be discussed.

In most of the scientific publications, as an attempt to identify the thematic areas within which an article falls and also to signify the specific contribution of that article, authors are prompted to list some suitable keywords. This can be treated as something that strengthens the conjecture that 'keywords can be used to represent thematic areas', though this fact does not provide establishing evidence. Apart from author keywords or author provided keywords that will be represented throughout this article as K(A), keywords extracted from article text are also used for various kinds of explorations. Several studies under the umbrella of 'text mining' extracts important terms from title, abstract and other parts of the article for various purposes, including the exploration of knowledge structure/organization within research fields. Indexing platforms or databases itself sometimes have inbuilt mechaanisms to extract keywords. Keyword plus terms, that will be denoted as K(P) throughout this article, are one such kind of extracted keywords provided by WoS (from the title of referencaes associated with an article). Investigation by Zhang et al. (2016) indicated that both author keywords (K(A)) and keyword plus (K(P)) are equally effective for investigation of knowledge structure within scientific fields. Following this, as a starting point, we designated both author keywords and keyword plus as representative terms for thematic areas.

We developed our framework for determining core competency and identifying thematic research strength of institutions by using the keyword information in articles and its subsequent processing using NLP and Network approach. The Fig. 1 shows the schematic diagram of the framework developed.

The step-by-step explanation of the procedure is as follows:

Step 1: Data collection

Institution-wise data for a country can be collected (selection of top institutions can be done based on some well-defined criteria) for the discipline under consideration. Details of data collection for the present case of Indian institutions are given in the Sect. 3.

Step 2: Data pre-processing

Data pre-processing is one of the most important steps because the accuracy of results and analysis depends greatly on the pre-processed data. Input data file that comes from major databases like WoS (Web of Science) consists of information against each document that include author keywords (column with field tag 'DE' of the WoS file) and keyword plus (column with field tag 'ID'). As we have stated, our starting point will be these keywords. However, a major problem associated with these keywords is that some of the keywords might not be suitable for representing a thematic area because (i) they may be too generic to the context of the discipline or may be at much coarser level than what is expected to be the level of a thematic area, or (ii) they may be at much finer level than what is expected to be the level of a thematic area. However, it is very difficult to know or define the exact level so that a keyword can qualify as a thematic area. Therefore, we had to explore and identify suitable NLP (Natural Language Processing) techniques to address and attempt to reduce the problems with the usage of raw set of author keywords and keyword plus. Thus, a suitable combination of NLP techniques that can be termed as 'NLP module' forms the crux of Data pre-processing step. This multi-layered NLP module is discussed next.

NLP module: In this step, we take the keywords and perform text pre-processing tasks. The author keyword field and keyword plus in the dataset has many ambiguities. Apart from these, plural and singular versions of some terms are also found. We have to replace these ambiguous keywords with more suitable ones (to represent thematic areas) and

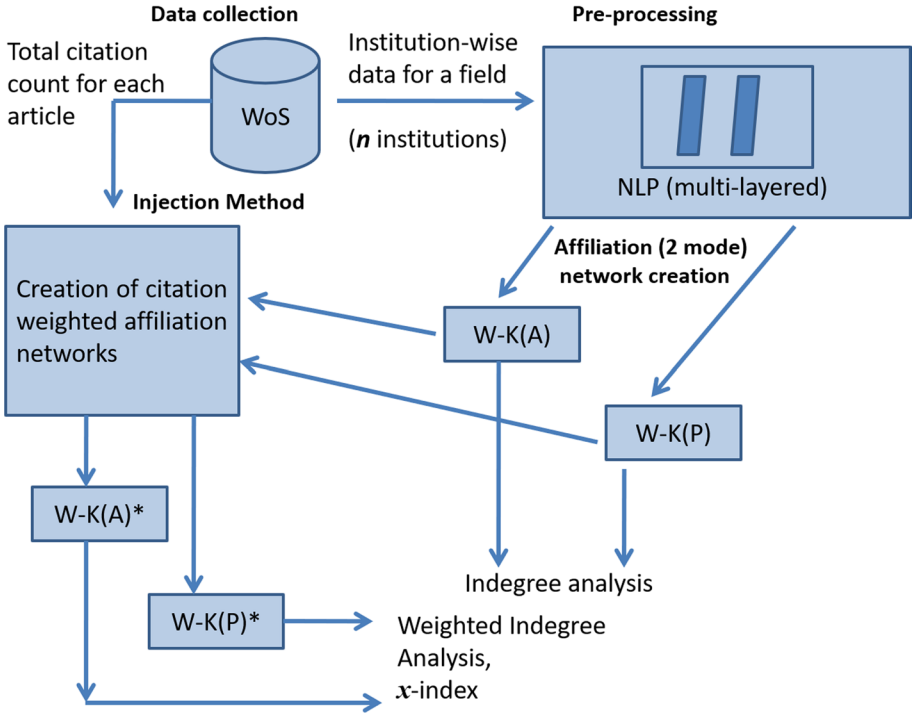

Fig. 1 The schematic diagram of the framework

replace the plural words with its singular counter parts. Thus, with the help of this multi-layered NLP module, we can obtain a reduced set of keywords, which is not only more representative (and hence improves accuracy) but also helps in reducing the computational costs. The schematic representation of the NLP module is given in Fig. 2.

As shown in the figure, a Word2Vec model is used to pre-process the keyphrases and identify similar keyphrases. The replacement or rephrasing of ambiguous keywords (i.e.,

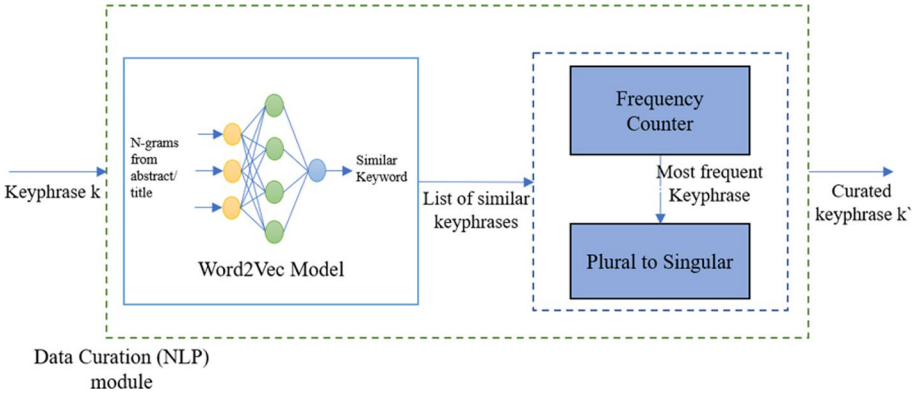

Fig. 2 Multi-layered NLP module

the function of first layer) is performed with the help of this word-embedding model. The model was trained on a subset of the Semantic Scholar Open Research Corpus dataset (Ammar et al., 2018). Only English articles were used from the whole dataset for the training process. From this step, we take only the most frequently occurring among the most similar keyphrases for each case, which helps to reduce the total number of keywords.

Initially we created a vocab (vocabulary) of all the author-provided keywords within the dataset. This vocab is then used to retrieve and store the top-5 most similar keywords for each of the keyword, using the trained Word2Vec model. Then, for each keyword and its similar keywords, we replace the keyword in the vocab with the most frequently occurring keyword (from the set of original keywords and similar keywords). Further, for the second layer that is supposed to replace the plural words with its singular versions, Levenshtein distance (Levenshtein, 1966) was used. From each pair of keywords with edit/ Levenshtein distance of 1, we have checked whether the last character of the phrase is 's'. If it is found so, the other phrase in that pair can be taken as the singular form and every occurrence of the plural term in the pair can be replaced by the singular term. A simple lemmatization would not suffice for this task since we had several multi-term keywords within our data which would be lost otherwise. A brief illustration on transformation of raw keywords to processed keywords by the NLP module is shown below:

Original Word	**Word after Pre-processing**
schemes	scheme
theorems	theorem
orientation	position
technology	application

Let K(A) and K(P) denote the NLP processed set of author keywords and keyword plus, respectively. Next important task is the identification of strength of an institution in each thematic area associated with it. This is discussed next.

Identification of strength of an institution in each thematic area requires mapping of the publications of the institutions to the keywords that designate thematic areas and computation of suitable scores that indicate the thematic strength. The network-based framework introduced by Lathabai et al. (2017) offers the provision to this mapping and computation of different kinds of scores (depending on the attribute used as input) through the procedure termed as *Injection methodology*. As citation is a major proxy for the impact of research, we use total citation counts for the Injection and computation of thematic strength. However, it is to be noted that the framework is surely capable of taking in different kinds of attributes as input and compute different kinds of scores. For instance, the usage of altmetric scores for injection can lead to the computation of early impact of thematic areas or thematic strength for visibility or popularity. However, as we are currently focusing on the determination of existing thematic strength, such explorations are reserved as future endeavors. As network creation is the key step (i.e., mapping of thematic areas to publications) that can serve as base for computation of scores to indicate thematic strengths, it is discussed next.

Step 3 Network creation

Before discussing network creation, some useful definition of networks and different kinds of networks are discussed.

A network is a structure $N=(V, L, F, W)$, that consists of a set of entities called vertices (V) linked together by a set of links L, so that $L \subseteq V \times V$. F is the vertex value functions or

properties. L can have both directed or undirected links (Batagelj, 2012). Set of undirected links are called edges (E) and set of directed links are called arcs (A). F is the vertex value functions or properties. Networks can be weighted or unweighted depending on the presence of link weights or vertex weights. A weighted network consists of a set of weights, W: L → ℝ or W: V → ℝ or both. For unweighted networks, all the links (and vertices) will carry unit weights. Affiliation networks or 2-mode networks (essentially bipartite) form the crux of this approach. These are discussed next.

Networks can be unimodal and multi-modal depending on the modalities (or the kinds of vertices they possess). A network is said to be unimodal if V is comprised only of one kind of vertices/entities (eg., article citation networks, patent citation networks, etc.,). A network is said to be multi-modal if it consists of more than one kind of vertices (eg; employee-firm networks, transportation networks where mode of transport is one kind of vertices and routes are another kind). Simplest one among the multi-modal networks is the 2-mode networks or affiliation networks. It consists of two types of vertices and the links usually exist between the two kinds of vertices and not among the same kind, making them bipartite. Formal definition of 2-mode networks is given by Batagelj (2012); Lathabai et al. (2017). Following this, definition of the Work-Keyword affiliation network (that represents the thematic area mapping of each publication) can be:

W-K network or the affiliation network of works and keywords, is a structure W-K = (W, K, L, P) where W is the set of works that forms the first mode and K is the set of keywords that forms the second mode and L is the set of arcs which originate from W and terminate at K, P is the weight of arcs. W-K affiliation network can be represented by:

$$P = \left[p_{wk}\right]$$

where,

$$p_{wk} = \begin{cases} p(w,k), if\ w,k \in L \\ 0 \qquad , otherwise \end{cases}$$

For an unweighted W-K affiliation network, default value of $p_{wk} = 1$, if there is an arc from work w to keyword k. For weighted affiliation networks, p_{wk} will take real values depending on the nature of mechanism that generated weights.

Depending on whCether author keywords or keyword plus are used, we can have two affiliation networks W-K(A) and W-K(P) respectively. We need unweighted affiliation networks as well as weighted affiliation networks with respect to K(A) and K(P) for the reason that will be mentioned below. Therefore, major processes in network creation step are given below:

(a) *Affiliation network creation*

Affiliation networks such as W-K(A) and W-K(P) networks are created from the processed data output of step 2 using the network creation module. Affiliation networks created here are directed and unweighted. For each institution, W-K(A) and W-K(P) networks are to be created. For this, the raw keywords in fields with tags DE and ID found as nested relations with respect to the publications in the field with tag UT should be processed to obtain the 1NF form. Now, the resulting edge list data structure can be used for the creation of affiliation networks either directly or with the help of auxiliary processing using popular network analysis and visualization packages such as Gephi (Bastian, Heymann & Jacomy, 2009) and Pajek (Batagelj & Mrvar, 1998). As we

require the creation of affiliation networks for all the institutions, we have extensively used the pajek library and networkx (graph-based) library provided in Python.

Weighted affiliation networks are created with the intention to compute scores of importance or scores that indicate performance of an institution with respect to certain attributes. As we said earlier, due to some limitations we demonstrate the framework using the attribute 'total (raw) citation' of publications. Therefore, weighted affiliation networks we create will be citation-weighted affiliation networks.

(b) *Citation-weighted affiliation network creation through Injection*

Injection methodology introduced by Lathabai (2017) can be used to create citation weighted affiliation networks. Information about total number of citations earned by publications (in WoS and some major databases) can be found in the data field designated with field tag 'z9' in WoS file. These values corresponding to each publication are injected to the W-K(A) and W-K(P) networks as arc weights and weighted affiliation networks designated as W-K(A)* and W-K(P)* can be created in this step as shown in Fig. 3

The injection methodology can be achieved in the package Pajek. But for achieving this for 195 institutions in a time-effective manner, we have used Pajek library and networkx library.

For illustration, the weighted W-K(A) network (injected with citation weights) of IIT Kharagpur, a pioneer institution in India, is shown in Fig. 4.

Unweighted W-K(A) affiliation network of IIT Kharagpur consists of 5734 vertices (1,205 works in 1st mode and 4529 keywords in 2nd mode) and 6131 arcs. Citations earned by the publications are injected as weight of links of affiliations from publications to the keywords through the injection process. Citation values are available from

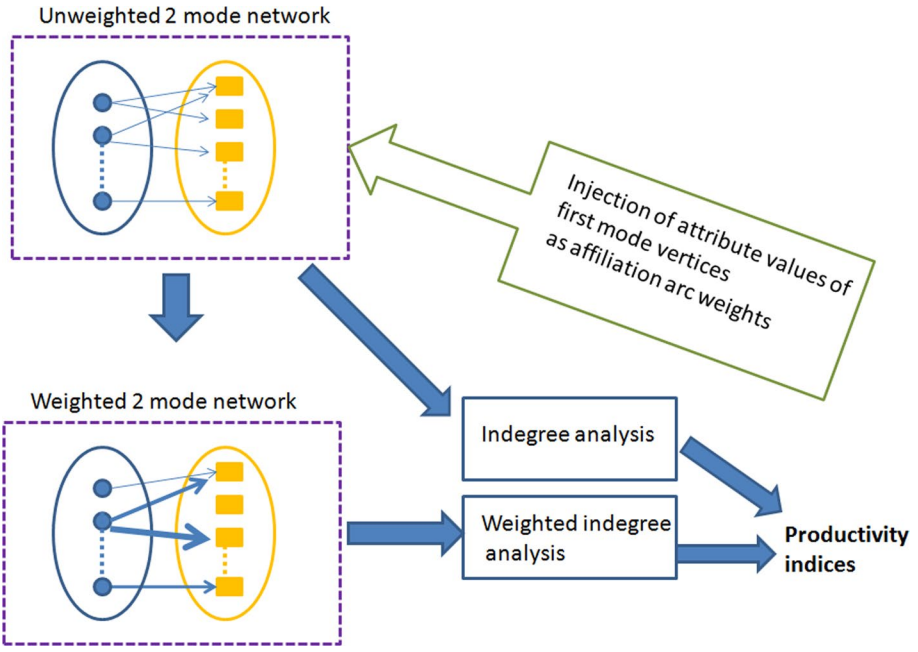

Fig. 3 The Injection methodology (introduced by Lathabai et al. (2017))

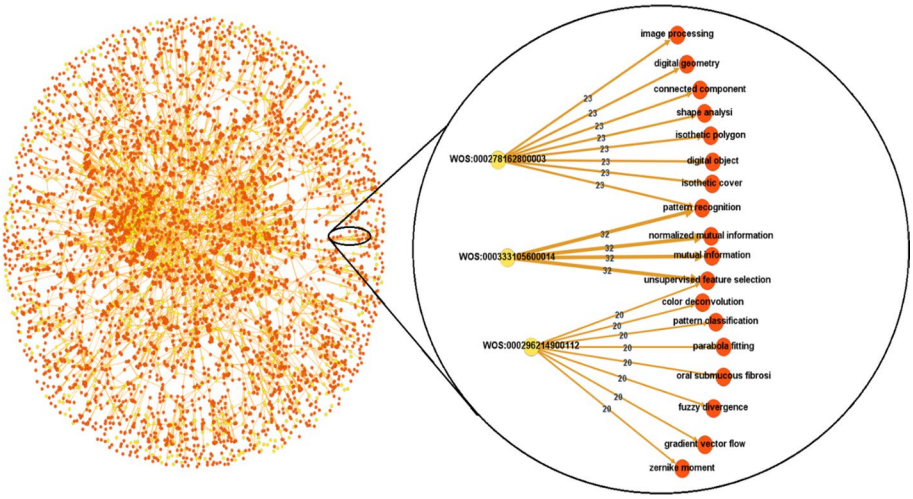

Fig. 4 The citation weighted W-K(A) network of IIT Kharagpur highlighting a subnetwork

the downloaded WoS data file in the data field with field tag 'Z9'. Weights are injected after matching the labels of 1 st mode vertices and publication IDs in the field with tag 'UT' in the WoS file. This creates the citation weighted affiliation network W-K(A)* which is shown in Fig. 4. The yellow vertices are the works (first mode) and red vertices (second mode) are the keywords. A small subnetwork containing 21 vertices (3 publications and 18 keywords) is highlighted. Three publications WOS:00,027,816,280,003, WOS:000,333,105,600,014 and WOS:000,296,214,900,112 shown in Fig. 4 have received 23, 32 and 20 citations respectively. These are injected to the network as arc weights and hence all the arcs originating from WOS:00,027,816,280,003 and directed towards keywords (in second mode) are assigned 23 as weight. In this way all the arcs shown in the subnetwork and rest of the network are assigned weights to obtain W-K(A)*. The network and subnetwork visualizations shown in Fig. 4 are achieved with the package Gephi.

Step 4: Network analysis.

Injection methodology introduced by Lathabai (2017) gives two methods for analysis of affiliation networks- indegree analysis of unweighted affiliation network and weighted indegree analysis of weighted affiliation networks. These are described below and the injected methodology that forms step 3 and 4 is depicted in Fig. 3. Apart from this, some more analysis methods (shown in Fig. 1) will also be discussed in this step.

(a) *Indegree analysis for number of publications*

In a directed network, indegree of a vertex indicates the number of incoming arcs to that vertex. In case of 2-mode affiliation network such as W-K(A), for author keywords, the indegree of a keyword gives the number of publications in which the author keyword appears. Indegree of works or publications in W-K(A) network will be zero. Therefore, as we deal with institutional data, indegree of author keyword gives the number of publications by an institution in a particular thematic area (designated by the author keyword). Similar is the case of keyword plus too.

(b) *Weighted indegree analysis for computation of thematic strengths*

In a weighted directed network, weighted indegree of a vertex indicates the sum of weights of incoming arcs to that vertex. For citation weighted W-K(A)* network obtained in step 3, the weighted indegree of keyword gives the total number of citations earned by that thematic area represented by the keyword due to the publications in which that particular keyword appears. Weighted indegree of publications will be zero. Since we are dealing with institutional data, weighted indegree of author keyword gives the number of citations earned by an institution in a particular thematic area (designated by the author keyword). Same procedure can be used for keyword plus too.

For instance, the thematic strength of IIT Kharagpur in the area 'unsupervised feature selection' can be computed as the weighted indegree of the area, which is found to be 52 as two publications with citations 32 and 20 are mapped to the area 'unsupervised feature selection'. Most of the network analysis and visualization packages comes with the provision for indegree and weighted indegree analysis. However, for conducting these for all the institutions we have used network library.

Now, a mechanism that could compute the performance of the institution with respect to the discipline from the performance in thematic areas can achieve the stated objective. One of the sensible ways to do this is to identify the core competency thematic areas of an institution and consider the consolidated value that defines the core competency areas of the institution as an indicator of the performance of the institution in that discipline. Inspired by the fact that *h*-index (Hirsch, 2005) and *h*-type indices such as *g*-index (Egghe, 2006) partition the profile (that consists of all the publications and the respective citations received by these publications) of an actor into core (*h*-core, *g*-core, etc.,) and tail (*h*-tail, *g*-tail, etc.,), we devise a mechanism to identify the core competency thematic areas and potential core competency thematic areas of an institution by adopting these indices. As the core competency areas and potential core competency areas may also signify the areas in which an institution can be regarded to be of 'extremely high' and 'high' expertise, the indicators that are designed to identify these areas can be termed as *expertise indices*. Two major expertise indices that can be designed are *x*-index (Lathabai, Nandy & Singh, 2021) and *x(g)*-index using the underlying principle of *h*-index and *g*-index respectively. These can be defined as:

***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 with thematic strengths of at least *x*. Here the thematic strengths are computed as total citation scores or altmetric scores received for those areas. These *x* areas that form the *x*-core can be treated as the core competency areas of the institution.

***x(g)*-index**: An institution is supposed to have an *x(g)*-index value of *x(g)* if it has published papers in at least *x(g)* thematic areas such that the average thematic strength from these areas amounts at least to *x(g)*. These *x(g)* areas (including top *x* areas and next (*x(g)*-*x*) areas), that form the *x(g)*-core, can be treated as a collection of core competency areas and potential core competency areas. The term 'potential core competency area' is used because these areas can become core competency of an institution in immediate future or later depending upon the level of emphasis the institution is going to place in such areas.

In this work, we also attempt to compare the above expertise indices to determine the better one for achieving the objective of this research. Now, the computational information of these indicators that forms the basis of third analysis in step 4 is given below:

(c) *Computation of x-core and x(g)-core from weighted indegree results*

From the sorted profile of an institution (sorted by weighted indegrees of keywords), x-index of an institution can be computed. x-index satisfies the following condition.

$x = \{max_i : weighted\ in\ degree\ of\ keyword\ at\ position\ i \geq i\}$.

$x(g)$-index satisfies the following condition.

$x(g) = \{max_i : \sum_i weighted\ in\ degree\ of\ keyword\ at\ position\ i \geq i^2\}$.where, i is an arbitrary rank/position of sorted keywords (sorted by weighted indegrees).

Once the thematic strengths (weighted indegrees) of each thematic areas are known, simple python code using libraries such as NumPy can be used for computation of expertise indices.

Computation of x and $x(g)$ indices is demonstrated in Table 1 and Fig. 5 using the case of the institution 'Dr. Mahalingam college of engineering and technology (DMCET), Pollachi, Tamil Nadu'.

From Table 1 and Fig. 5 found to be 11 and 18 respectively. This implies that 11 thematic areas (top 11 areas in Table 1) can be regarded as core competency or core expertise areas of DMCET. Since $x(g)$-index is found to be 18, the seven thematic areas that are found in $x(g)$-core apart from the thematic areas in x-core can be treated as potential core competency areas.

An institution can have different expertise index values (i.e., different x and $x(g)$ indices) in different disciplines. Thus, policymakers or funding agencies that intend to provide 'thrust area' based funding can rely on x-index and/or $x(g)$-index to identify the deserving institution in that thrust area rather than considering the overall performance indicators like h-index, publication and citation counts, etc. Since, the proposed index takes inputs

Table 1 Demonstration of computation of x and $x(g)$ indices of the institution DMCET

Rank	Thematic areas	Strength (weighted indegree)	Squared rank	Cumulative strength
1	Support vector machine	38	1	38
2	Wavelet transformation	35	4	73
3	Eeg classification	35	9	108
4	Machine learning	32	16	140
5	Epileptic seizures	32	25	172
6	Feature selection	17	36	189
7	Cascaded multilevel inverter	15	49	204
8	Anfis	15	64	219
9	Control voltage	15	81	234
10	Grid voltage	15	100	249
x-index = 11	pv	15	121	264
12	Image texture	11	144	275
13	Matrix algebra	11	169	286
14	Feature extraction	11	196	297
15	Image segmentation	11	225	308
16	Active contour segmentation technique	10	256	318
17	Ultrasound images	10	289	328
$x(g)$-index = 18	Segmented region	10	324	338
19	Wavelet features	10	361	348

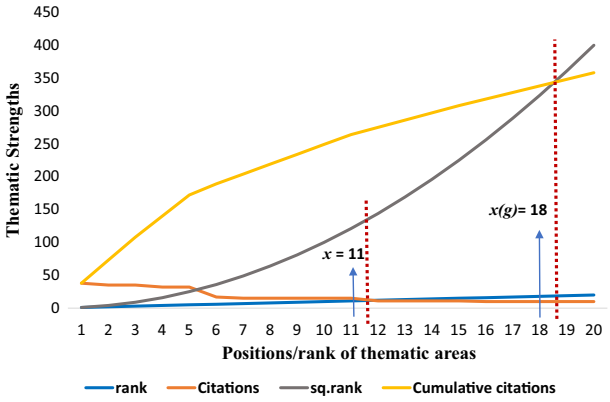

Fig. 5 Graphical demonstration of computation of expertise indices

from both, the core competency and thematic research strength, it in a way combines both breadth and depth of research of an institution.

The principle of expertise cores such as *x*-core and *x(g)*-core can be further extended to have nested expertise cores. Nested *x*-index and *x(g)*-index of a country can be used to find its competency in a discipline. The nested *x*-index of a country can be defined in the following way:

A country is supposed to have a nested *x*-index of *x* if it has at least *x* institutions with *x*-index values of *x*.

A country is supposed to have a nested *x(g)*-index of *x(g)* if it has at least *x(g)* institutions whose *x*-index values averages to *x(g)*.

Similar to the case of expertise indices of institutions being capable of determining core competency areas and potential core competency areas an institution in a field, the nested expertise indices can determine the core competent institutions and potential core competent areas of a country in a field. Potential core competent institutions mean these may rise and join the ranks of core competent institutions either in the immediate future or later. Notion of nested expertise indices can be used for comparison of countries. However, such an exercise involves extensive amount of data and computations and is beyond the scope of this work, as we are mainly focused on institutional performance assessment. The nested indices-based comparisons may be attempted as a future endeavor. Now, we demonstrate the framework using some of the results.

Results

For each of the 195 Indian institutions selected, after pre-processing using multi-layered NLP module, we have created W-K(A) and W-K(P) networks following steps 2 and 3.

It is converted to W-K(A)* network after injection of citation weights. W-K(A) network (after processing with multi-layered NLP), when subjected to indegree analysis provides the number (frequency/count) of publications within which a keyword appears. In case of IIT-Kharagpur, ‘genetic algorithm’ is the most frequent keyword (appears in 37 publications) and ‘security’ is the second most frequent keyword (its frequency is 36). W-K(A)* network, when subjected to weighted indegree analysis provides the total number of citations received by each keyword (via the publications in which they appear). Top 25

thematic areas of IIT Kharagpur according to author keywords with respect to (i) number of publications and (ii) number of citations is given in Table 2.

Though 'genetic algorithm' is the top-most frequent keyword that appears in IIT Kharagpur's publications, it is ranked as 7th most cited keyword. 'wireless sensor network' is in the first position in terms of citations. *x*-index of an institution, that indicates the largest number of '*x*' keywords that has received at least '*x*' citations, can be computed in *h*-like fashion. In case of IIT Kharagpur, 116 is the expertise index (according to author keywords (K(A))). That means it has at least 116 author keywords that has received at least 116 citations. Similarly, W-K(P) network and W-K(P)* network of IIT Kharagpur and other institutions are also subjected to Indegree analysis and weighted indegree analysis respectively. *x*-index of IIT Kharagpur according to keyword plus (K(P)) is found to be 112.

The value of *x*-index for different institutions can be used to rank them according to expertise. Higher *x*-index value of an institution indicates that it has higher expertise of working on different themes of a subject (Computer Science in this case). Thus, in a sense the x-index value represents the breadth of research capability of an institution, represented as core competency. Top 10 institutions for *x*-indices value according to author keywords (K(A)) and keyword plus (K(P)) are shown in Tables 3 and 4, respectively. Here, IIT Kharagpur is at the top for author keywords (K(A)) and Thapar institute of engineering and technology is at top in calculation for keyword plus (K(P)). The *x(g)*-index of different institutions are also computed. Top 10 institutions for *x(g)*-indices according to author keywords and keyword plus are also shown in Tables 3 and 4, respectively. Here, Thapar institute of engineering and technology is the top institution according to author keywords (with *x(g)*-index of 191) and keyword plus (with *x(g)*-index of 209). Both, Thapar institute and IIT Kharagpur (with *x(g)*-index of 209) shares the top most position in terms of *x(g)*-index according to keyword plus.

The *x*-index indicates the core competency areas of an institution. *x(g)*-index indicates the core competency as well as potential core competency areas. The thematic areas that are found in *x(g)*-core below the top *x* areas, i.e., thematic areas found at positions from $x+1$ to *x(g)* are the potential core competency areas. So, if two institutions *i* and *j* have x_i and x_j as their *x*-indices and $x(g)_i$ and $x(g)_j$ as their *x(g)*-indices, then if we observe $x(g)_i > x(g)_j$ but

Table 2 Top 25 thematic areas of research of IIT Kharagpur in terms of number of publications and thematic strength (according to Author keywords)

S. No	Keyword (Au)	Publication count	S. No	Keywords (Au)	Citation count
1	Genetic algorithm	37	1	wireless sensor network	1007
2	Security	36	2	Security	895
3	Particle swarm optimization	31	3	Internet of thing	709
4	Wireless sensor network	30	4	Cloud computing	641
5	Authentication	21	5	Avispa	623
6	Neural network	20	6	Authentication	581
7	Learning automatum	17	7	Genetic algorithm	546
8	Cloud computing	16	8	Learning automatum	499
9	Simulation	15	9	Fog computing	484
10	Network-on-chip	15	10	Particle swarm optimization	482

Table 3 Top 10 Indian HEIs by *x* and *x(g)*-indices (according to author keywords)

S No.	Name of institution	*x*-index (Au)	S No.	Name of institution	*x(g)*-index (Au)
1	Thapar institute of engineering and technology	116	1	Thapar institute of engineering and technology	191
2	Indian institute of technology IIT Kharagpur	115	2	Indian institute of technology IIT Delhi	188
3	Indian institute of technology IIT Delhi	104	3	Indian institute of technology IIT Kharagpur	181
4	Indian statistical institute Kolkata	101	4	Indian statistical institute kolkata	180
5	Indian institute of technology IIT Roorkee	95	5	Indian institute of technologY IIT Roorkee	157
6	Vellore institute of technology	84	6	Vellore institute of technology	141
7	Indian institute of technology IIT Kanpur	77	7	International institute of information technology hyderabad	140
8	Indian institute of science IISC Bangalore	74	8	Anna university	129
9	National institute of technology Rourkela	70	9	Indian institute of science IISC Bangalore	125
10	Anna university	65	10	Indian institute of technology IIT Kanpur	124

Table 4 TOP 10 Indian HEIs by *x* and *x(g)*-indices (according to keyword plus)

S. No.	Name of institution	*x*-index (KP)	S. No.	Name of institution	*x(g)*-index (KP)
1	Indian institute of technology IIT Kharagpur	113	1	Indian institute of technology IIT Kharagpur	209
2	Thapar institute of engineering and technology	111	2	Thapar institute of engineering and technology	209
3	Indian statistical institute Kolkata	100	3	Indian statistical institute Kolkata	196
4	Indian institute of technology IIT Roorkee	96	4	Indian institute of technology IIT Delhi	180
5	Indian institute of technology IIT Delhi	93	5	Indian institute of technology IIT Roorkee	172
6	Vellore institute of technology	92	6	Vellore institute of technology	151
7	Indian institute of technology IIT Kanpur	77	7	Indian institute of technology IIT Indore	130
8	Indian institute of science IISC Bangalore	69	8	Indian institute of technology IIT Kanpur	129
9	Indian institute of technology IIT Madras	66	9	International institute of information technology Hyderabad	128
10	Indian institute of technology IIT Indore	66	10	JNU New Delhi	127

$x_i = x_j$, this can indicate that though institutions i and j are having x core competency areas (areas can differ), institution i have more potential core competency areas than j and therefore i may have more chance to overtake j in terms of x-index value in immediate future or later, than vice versa. Thus, while x-index reflects the present expertise, $x(g)$-index offers room for speculation. From Tables 3 and 4, Thapar institute has 116 or 111 core competency areas and 75 (191–116) or 98 (209–111) potential core competency areas. In case of IIT Kharagpur, the number of core competency areas are 115 or 113 and the number of potential core competency areas are 66 (181–115) and 96 (209–113). Thus, in terms of both author keywords (K(A)) as well as keyword plus (K(P)), Thapar institute has more potential core competency areas than IIT Kharagpur. Therefore, Thapar institute has to potential to overtake IIT Kharagpur or become equal to IIT Kharagpur in terms of the number of core competency areas in future.

On the other hand, though a difference in values is visible for the x-indices and $x(g)$-indices obtained with author keywords and keyword plus, there is no significant change in the relative ranking/ordering of institutions. This can be verified using the Spearman's rank correlation analysis. In the case of x-index, the correlation co-efficient (Spearman's ρ) between the x values computed from author keywords (K(A)) and keyword plus (K(P)) is found to be 0.916. Therefore, we can rule out the chance of existence of high ranking of an institution with author keyword and low ranking of an institution with keyword plus and vice versa. The scatter plot shown in Fig. 6 illustrates this.

The four quadrants shown in the Fig. 6 are determined in the following way:

Quadrant 1 (Low DE- Low ID)

$$Low\,DE\,:\,x-index\,(DE)\,<\,0.5\times max\{x-index\,(DE)\}$$

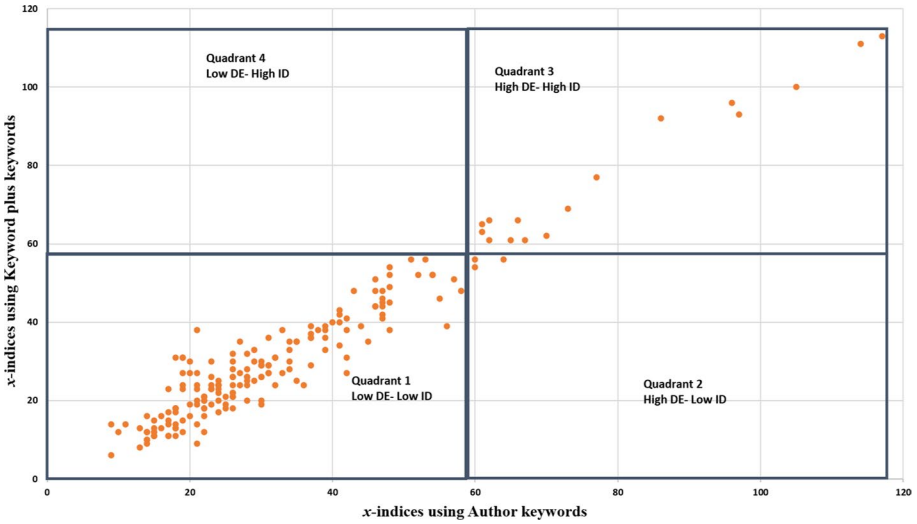

Fig. 6 The scatter plot of x-indices computed using Author keywords and Keyword plus

$$Low\,ID\,:\,x-index\,(ID)\,<\,0.5\times max\{x-index\,(ID)\}$$

Quadrant 2 (High DE- Low ID)

$$High\,DE\,:\,x-index\,(DE)\,\geq\,0.5\times max\{x-index\,(DE)\}$$

$$Low\,ID\,:\,x-index\,(ID)\,<\,0.5\times max\{x-index\,(ID)\}$$

Quadrant 3 (High DE- High ID)

$$High\,DE\,:\,x-index\,(DE)\,\geq\,0.5\times max\{x-index\,(DE)\}$$

$$High\,ID\,:\,x-index\,(ID)\,\geq\,0.5\times max\{x-index\,(ID)\}$$

Quadrant 4 (Low DE- High ID)

$$Low\,DE\,:\,x-index\,(DE)\,<\,0.5\times max\{x-index\,(DE)\}$$

$$High\,ID\,:\,x-index\,(ID)\,\geq\,0.5\times max\{x-index\,(ID)\}$$

where max{x-index (DE)} = 116 and max{x-index (ID)} = 113.

So, the boundaries of quadrants are chosen at 58 of X-axis and 56.5 of Y-axis.

As we can see, almost all the institutions belong to Low DE- Low ID and High DE- High ID quadrants. Only three institutions lie in quadrant 2, that too near the boundary of quadrant 3.

Similar is the case of *x(g)*-indices as the Spearman's rank correlation is found to be 0.935.

Thus, though there is a limitation in thematic area designation using author keywords (K(A)), keyword plus (K(P)) or any other extraction methods that hinders computation of results with absolute accuracy, the overall picture of performance of institutions can be obtained using such exercises, unless the keyword extraction method is highly ineffective.

As our framework is capable of identifying the performance of institutions with respect to thematic areas, it can be used to determine which institutions are better in a specific thematic area. Thus, the framework can be used at the lowest level of granularity of a given keyword as well. For instance, in the thematic areas 'Wireless Sensor Network' and 'Data Mining', top 15 Indian institutions are given in Table 5 using author keywords (K(A)). The institutions are ranked by citations they received in the given thematic area. An alternative could be to rank the institutions by their performance in terms of any other relevant indicators in the given thematic area, or may be simply by number of papers in the given thematic area. Similar exercise can be done for keyword plus (K(P)) also.

It can be observed that the two rankings of institutions (one on core competency and other on thematic strength) are quite different indicating that different institutions have different research strengths in different thematic areas. In fact, the rankings at the two levels of granularity can be used together for different kinds of assessment. While, the x-index,

Table 5 Top 15 Indian institutions in the area 'Wireless sensor network' and 'Data mining' by citation counts (according to author keywords)

Rank (Au)	Name of institution (Wireless sensor network)	Citation count	Rank (Au)	Name of institution (Data mining)	Citation count
1	IIT (ISM) Dhanbad	1291	1	ISI Kolkata	219
2	IIT Kharagpur	1056	2	IIT Kharagpur	140
3	VIT	837	3	Gautam Buddha University	138
4	Chaudhary Charan Singh University	795	4	Malaviya national institute of technology Jaipur	138
5	International institute of information technology, Hyderabad	536	5	Guru Ghasidas Vishwavidyalaya	123
6	Anna University	458	6	JNU New Delhi	119
7	Thapar institute of engineering and technology	449	7	IIT Indore	119
8	ISI Kolkata	255	8	IIM Ahmedabad	114
9	NIT Rourkela	244	9	Anna University	96
10	IISC Bangalore	228	10	NIT Silchar	91
11	I K Gujral Punjab technical University	216	11	Indian institute of engineering science technology, Shibpur	86
12	Indian institute of engineering science technology, Shibpur	205	12	International institute of information technology Hyderabad	85
13	DR B C Roy Engineering college	195	13	Thapar institute of engineering and technology	77
14	JNU New Delhi	174	14	Annamalai University	73
15	IIT Delhi	167	15	NIT Kurukshetra	57

representing the core competency of an institution in a given subject, can be seen as a measure of breadth of research capability of the institution; the thematic research strength of the institution in a specific theme can be seen as depth of research capability of the institution in that specific theme. Thus, in a way, both can be used together, with different weights assigned to each. This capability of our framework will, therefore, be useful for performance-based funding in 'thrust areas/national priority areas' and also for various other purposes. Suppose if the funding agency is assessing institutions for funding depending on their performance in more than one thrust-area (areas can be closely related to each other), then ranking based on thematic strengths in these thrust-areas can be helpful. Moreover, a base research strength of the institution can be determined by the x-index value computation.

Discussion

The paper presents a framework for identifying the core competency of institutions in a given subject and asses their research performance in different thematic areas of that subject. The data for Indian institutions in Computer Science is used to demonstrate the working of the framework. The framework provides for an expertise index (denoted as *x*-index) to represent the research capability of an institution in a given subject area. Higher *x*-index value denotes higher research capability of the institution in different thematic areas of the given subject. Similar to *x*-index, another expertise index namely the *x(g)*-index values are also computed. This enables the identification of potential core competency areas. An institution will have (*x(g)*-*x*) potential core competency areas. Thus, if two institutions have the same or almost equal *x*-index values, but the *x(g)* value of one of them is significantly higher than the other, then the number of potential core competency areas (*x(g)*-*x*) of the former will be greater than the latter. This can indicate that in the immediate future or sometime later, the institution with high *x(g)* is more likely to have more core competency areas (i.e., more *x*-index) than the other. This hints at the predictive capability of our framework. However, to validate this predictive capability, a temporal analysis has to be conducted. This is one of the possible enticing research explorations related to this framework. Further, in addition to identifying the core competency of institutions and potential core competency areas (measured by their *expertise indices*), the framework also allows the institutions to be ranked in their research performance in a specific thematic area of the given subject, according to thematic strengths. One may decide to rank institutions either in a specific thematic area or a set of 2–3 (or more) thematic areas combined, depending on the need, especially for evaluations for thrust-area funding. For instance, if a funding agency is looking for best institution for funding based on their performance in more than one thrust-area, then our framework can be used to determine thematic strengths of competing institutions in those thrust-areas and rank institutions based on that. The institution that has been ranked high in most of the areas can be chosen for funding or given more preference/weightage in the overall decision process if the process involves other assessment methods too. At the same time, the *x*-index value can be used as a measure of core competency (and hence base research strength) of the institution in a subject, and can be combined with chosen thematic research strengths. Further, if the purpose is to develop centre of research excellence in a given subject area, the *x*-index value can be directly used for ranking. Thus, the framework can be useful for subject-area as well as thrust area-based research funding tasks of various types.

We intend to compare the rankings obtained with our framework to other popular rankings for checking whether there is an essential difference or novelty in our ranking. Our dataset has 195 Indian institutions and major ranking systems have very limited number of Indian institutions listed in their subject rankings for Computer Science. Therefore, we are using relative ranks of Indian institutions in these ranking systems and relative ranks of institutions according to expertise *(x* and *x(g))* indices for comparison. Spearman rank correlation is used for comparison. Firstly, the comparison for *x*-indices computed using author keywords is conducted. For ARWU, 7 institutions are common and Spearman's ρ= 0.571. For THE, 17 institutions are common and ρ is found to be 0.678. For QS, 16 institutions are common and ρ obtained is 0.496. In case of CWTS Leiden, results are available only in a format where fields 'Mathematics' and 'Computer Science' are combined. For CWTS, 31 common institutions were found and ρ= 0.801. In case of keyword plus, Spearman's ρ values obtained for ARWU, THE, QS and CWTS are 0.393, 0.698, 0.568 and 0.766, respectively. In case of *x(g)*-indices, the Spearman's ρ for relative rankings of common institutions are 0.75, 0.665, 0.404 and 0.76 with respect to ARWU, THE, QS and CWTS rankings, according to author keywords. For keyword plus, ρ scores are 0.295, 0.659, 0.501 and 0.676 respectively for ARWU, THE, QS and CWTS rankings.

From the above results, it can be inferred that our ranking framework generates essentially different rankings from most of the existing ranking systems, as one would expect since we are applying the framework to measure the research capability of institutions in terms of their core competency. Though slightly higher correlation of expertise-based rankings is observed especially with CWTS ranking (where CWTS considers mainly the research publication data for ranking, though it includes Mathematics too with Computer Science), the limitations of CWTS Leiden for thematic area-based ranking (such as usage of broad subject categories, and under-representation of developing countries etc.), make the proposed framework a worthy choice for thematic area-based rankings. Our framework not only determines the core competency of institutions in a given subject but also allows identifying research strength of institutions in specific thematic areas. One may even combine multiple thematic areas together, to determine the research strength in the given multiple thematic area combine. Further, funding decisions can be based ranking on both, the core competency (indicating research breadth in a subject) and thematic research strength (indicating research depth in a given thematic area of the given subject).

As design of expertise indices is inspired from the *h*-type indices, it will also be interesting to check whether the expertise indices of institutions correlate well with the institutional *h*-type indices. By institutional *h*-type indices we mean the h-type index that is computed from the scientific publication profile of an institution. For instance, institutional *h*-index of an institution is the largest number of *h* publications (in the profile of an institution) that are cited at least *h* times. As *g*-index is a well-known *h*-type index, institutional *g*-index of an institution is the largest number of *g* publications whose citations averages at least to *g*. Correlation between *x*-index values of institutions for both author keywords (K(A)) and keyword plus (K(P)) are compared with institutional *h*-index values, and correlation between *x(g)*-index values of institutions for both author keywords (K(A)) and keyword plus (K(P)) are compared with institutional *g*-index values. Spearman's ρ values between institutional *h*-indices and *x*-indices for author keywords and keyword plus are 0.932 and 0.925, respectively. Correlation between institutional *g*-indices and *x(g)*-indices for author keywords and keyword plus are 0.97 and 0.943. Thus, high level of correlation is found between expertise indices and institutional *h*-type indices. Therefore, expertise-based indices are as useful as institutional *h*-type indices for assessment of institutional performance. However, there is an added advantage with expertise-based indices that perhaps

make the expertise-based assessment more useful for institutional performance assessment than *h*-type institutional indices. The *h*-type indices in one way or other represent the overall productivity of an institution, while expertise-based indices give an idea about the width or diversity of the expertise of an institution and thereby provides the information about the possible core competency areas of an institution. Further, it is possible to look into the results at different levels of granularity and thus determining the research strengths of institutions in one or more given specific themes. The two ranks can also be combined together in different ways for various applications.

We have used citation counts for creating "citation-weighted affiliation networks" and also for ranks based on thematic strengths. Though, other attributes (such as publication counts or altmetric counts) can be used for the ranking purpose, but one needs to keep in mind the inherent biases introduced due to use of raw citation counts. Citation counts are likely to vary not only across subjects but also within with different thematic areas of a particular discipline. For example, theoretical areas and applied areas will have different publication and citation patterns. Applied and emerging areas (themes) of computer science research get cited more frequently than the theoretical areas (themes). Thus, it would be desirable to normalize the values of performance-score drawn through "citation-weighted affiliation networks". Use of source normalized citation values may be one possible option that remains to be explored. Though, we are comparing research strengths of institutions only within a thematic area and not across them, but the expertise indices (x and x(g)) use the data for different kinds of thematic areas and hence may be impacted by citation variations. This issue of using normalized citation values at the level of thematic areas needs more systematic effort and is reserved for future exploration.

Conclusion

Motivated by the persisting gap related to the limitations of existing frameworks for ranking institutional performance for funding in thrust areas of research (thematic areas), we introduced a framework that considers the thematic area strengths of institution for ranking its performance within a subject/ field. The framework inherits all the merits of NLP processing and network approach for contextual productivity assessment. The framework is capable of computing the strengths of an institution in different thematic areas using attributes such as citations, altmetric scores, etc. In this work, the citation is used to determine thematic strengths. The framework is also capable of computing the core competency areas and potential core competency areas of an institution. Core competency thematic areas are determined using expertise indices namely *x*-index and *x(g)*-index, the indicators that works on the principle of *h*-index and *g*-index, respectively. While *x*-index indicates the expertise of an institution in terms of the level of richness or diversity of research themes in which an institution is really strong, *x(g)*-index handpicks the thematic areas in the research portfolio of an institution that have the potential to join the ranks of core competency areas in immediate future or later. Also, as discussed earlier, among the institutions that are having same number of core competency areas, those having a greater number of potential core competency areas can be speculated to develop more core competencies than others in immediate future or later. Thus, apart from providing the knowledge of current stature or position of an institution in a research field/subject, the information about its core competency and potential core competency areas can be vital for institution level policymakers to work towards improvement of the institutional performance.

On comparison with existing ranking systems, our framework is found to provide essentially different rankings that hints the novelty/non-obviousness of ranking. Upon computation of correlation of expertise indices with institutional h and g indices, expertise indices are found to have high correlation with h and g indices, confirming that expertise indices are useful like institutional h and g indices for institutional ranking. Our framework has an added advantage that it is able to measure both, the core competency of an institution in a subject and its research strengths in different thematic areas of the subject. The two rankings can be used in several ways. One may use performance rankings on a group of thematic areas together for thrust area-based funding decisions. The x-index values can also be combined with such an exercise to provide an overall research capability assessment of institutions, as added evidence. Similarly, a pure x-index based assessment can be used for funding decisions aimed at developing centres of research excellence in a given subject.

Limitations and future work

Our framework has certain limitations too. First, in its present form, it is dependent on 'author keywords' and 'keyword plus' for determination of thematic areas. Therefore, the accuracy of the framework is dependent on the ability of such keywords to represent thematic areas. 'Natural Language Processing' (NLP) module (with state-of-the-art NLP techniques) is incorporated in the pre-processing phase of the framework to ensure the accuracy of the framework. However, there is always scope for improvement and the design of the framework is flexible enough to accommodate more advanced NLP techniques when they materialize. For example, one may choose to use NLP after pre-processing as well. This may include clustering the pre-processed thematic keywords into more coarse groups and label the groups as a higher-level of research theme. Use of advanced clustering algorithms can thus further improve the applicability of the framework. The domain knowledge about the structure of the discipline can be further useful for this purpose.

Another limitation could be that the framework is demonstrated only for a given subject and data is drawn from a specific database (Web of Science). The robustness of the framework can be re-affirmed if it is executed for data from other scholarly databases such as Scopus, Dimensions, etc. and for certain other subjects. This extension is reserved as a further exploration. Availability of domain knowledge of the subject so chosen will be an important advantage for applying the framework and assessing its suitability.

Another possible limitation can be related to the choice of attributes for the determination of the thematic strength. For instance, when the citation is used as the attribute for injection and thereby for the computation of thematic strengths, such a computation may be prone to an alleged bias caused in favour of some thematic areas over others, as also indicated in the 'Discussion' section. For instance, some thematic areas in 'Computer Science' that belongs to applied research areas, especially the emerging ones tend to attract more citations than the ones that mostly deal with theoretical or fundamental research and this will cause concerns for comparison of institutions. As our framework never intends to compare the performance of institutions based on their thematic strengths in different areas, thematic strength computation might not be much affected by the usage of citation as the attribute. However, like the effect of field/area biases of citations in the h-index, the expertise indices are also susceptible to whether the thematic areas of an institution belonging to the realm of applied or basic research. To address this possible issue, some kinds of normalizations such as the ones at the sub-field/thematic area level can be considered to be executed before injection. Such

an exercise will help to answer the question "Does the thematic area level biases matters for computation of expertise indices?" and might solve it simultaneously. This is one of the roadmaps for further research that might improve/strengthen the current framework. Apart from this, the network-based framework provides enormous potential for unlocking more insights that can be useful for policy-making. In this work, the potential of network analysis is not fully explored. For instance, 'injection method' forms a major part of this framework in the same way it had been for the contextual productivity assessment framework by Lathabai et al. (2017). Therefore, one possible natural extension of the present work is the exploration of the use of the framework with any other suitable credit allocation scheme such as the fractional credit allocation to reduce or eliminate the inherent biases of the full credit allocation in the framework by Lathabai et al. (2017) to a substantial extent as discussed by George et al. (2020). Another possible related exploration would be creation and analysis of 'institutional coupling network of thematic areas' that could provide more insights about the possibility of progress or improvement in performance of core competency areas and potential core competency areas identified by the framework devised in this work. Thus, usage of more advanced network analysis techniques to extricate more insights from the framework is also considered for our future endeavors.

The ability to provide information about core competency and potential core competency areas that in turn allows the institution to develop policies for its own improvement along with funding agencies and other policymakers makes the expertise indices more utilitarian than institutional h and g indices. As already mentioned earlier, temporal analysis of dynamics of institutional performance using expertise indices can be useful for the validation of the predictive power of our framework. This is one of the possible research pursuits that can be vital for policymakers at the institutional and national levels. The conceptual framework of these indices can be extended beyond institutional performance assessments. For example, the concept of expertise indices like x-index and $x(g)$-index can be extended to compute the nested x-index and nested $x(g)$-index that can determine the competency of a country. For example, according to nested x-index, the expertise score of India is 44 (with author keywords) and 41 (with keyword plus) for the subject/discipline 'Computer Science'. According to nested $x(g)$-index, the expertise score of India is 82 (with author keywords) and 84 (with keyword plus) for the subject/discipline 'Computer Science'. This means that India has 41 or 44 core competent institutions and 41 or 40 potential core competent institutions in the subject 'Computer Science'. Similar computations can be done for other countries by processing their publication data in the chosen subject and the application of our framework can be extended to country level comparisons too. Analysis of strengths of institutions belonging to different countries in one or more specific thematic areas can also be done using the portion of our framework that computes thematic area strengths. Such explorations, however, are left to be carried out in future.

Acknowledgements 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 Performance of Indian Higher Education Institutions (HEIs)*', bearing Grant No. DST/NSTMIS/05/04/2019-20, for this work.

References

Abramo, G., & D'Angelo, C. A. (2014). Assessing national strengths and weaknesses in research fields. *Journal of Informetrics, 8*(3), 766–775.

Abramo, G., & D'Angelo, C. A. (2020). A novel methodology to assess the scientific standing of nations at field level. *Journal of Informetrics, 14*(1), 100986.

Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., ... & Etzioni, O. (2018). Construction of the literature graph in semantic scholar. arXiv preprint arXiv:1805.02262.

Anowar, F., Helal, M. A., Afroj, S., Sultana, S., Sarker, F., & Mamun, K. A. (2015). A critical review on world university ranking in terms of top four ranking systems. *Lecture Notes in Electrical Engineering, 312*, 559–566.

Bastian, M., Heymann, S., & Jacomy, M. (2009, March). Gephi: an open source software for exploring and manipulating networks. In Third international AAAI conference on weblogs and social media.

Basu, A., Banshal, S. K., Singhal, K., & Singh, V. K. (2016). Designing a Composite Index for research performance evaluation at the national or regional level: Ranking Central Universities in India. *Scientometrics, 107*(3), 1171–1193.

Batagelj, V. (2012). Social network analysis, large-scale. In A. Robert Meyers (Ed.), Computational complexity: Theory, techniques, and applications (pp. 2878–2897). New York: Springer.

Batagelj, V., & Mrvar, A. (1998). *Pajek-Program for Large Network Analysis. Connections, 21*(2), 47–57.

Beck, S., & Morrow, A. (2010). Canada's universities make the grade globally. The Globe And Mail. Retrieved from https://www.theglobeandmail.com/news/national/canadas-universities-make-the-grade-globally/article4326026/ on 10th Jan. 2021.

Billaut, J. C., Bouyssou, D., & Vincke, P. (2010). Should you believe in the Shanghai ranking? *Scientometrics, 84*(1), 237–263.

Bornmann, L., Stefaner, M., de Moya Anegón, F., & Mutz, R. (2014). What is the effect of country-specific characteristics on the research performance of scientific institutions? Using multi-level statistical models to rank and map universities and research-focused institutions worldwide. *Journal of Informetrics, 8*(3), 581–593.

Bowman, N. A., & Bastedo, M. N. (2011). Anchoring effects in world university rankings: Exploring biases in reputation scores. *Higher Education, 61*(4), 431–444.

Costas, R., & Bordons, M. (2007). The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics, 1*(3), 193–203.

De Boer, H., Jongbloed, B., Benneworth, P., Cremonini, L., Kolster, R., Kottmann, A., ... & Vossensteyn, H. (2015). Performance-based funding and performance agreements in fourteen higher education systems. Center for Higher Education Policy Studies.

Egghe, L. (2006). An improvement of the h-index: The g-index. *ISSI Newsletter, 2*(1), 8–9.

García, J. A., Rodriguez-Sánchez, R., Fdez-Valdivia, J., Torres-Salinas, D., & Herrera, F. (2012). Ranking of research output of universities on the basis of the multidimensional prestige of influential fields: Spanish universities as a case of study. *Scientometrics, 93*(3), 1081–1099.

George, S., Lathabai, H. H., Prabhakaran, T., & Changat, M. (2020). A framework towards bias-free contextual productivity assessment. *Scientometrics, 122*(1), 127–157.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *PNAS, 102*(46), 16569–16572.

Jeremic, V., Bulajic, M., Martic, M., & Radojicic, Z. (2011). A fresh approach to evaluating the academic ranking of world universities. *Scientometrics, 87*(3), 587–596.

Lathabai, H. H., Prabhakaran, T., & Changat, M. (2017). Contextual productivity assessment of authors and journals: A network scientometric approach. *Scientometrics, 110*(2), 711–737.

Lathabai, H. H., Nandy, Abhirup & Singh, V. K. (2021). Expertise-based institutional collaboration recommendation in different thematic areas. In proceedings of the 11th International Workshop on Bibliometric-enhanced Information Retrieval co-located with 43rd European Conference on Information Retrieval (ECIR 2021), Lucca, Italy (online only), April 1st, 2021.

Lazaridis, T. (2010). Ranking university departments using the mean h-index. *Scientometrics, 82*(2), 211–216.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *In Soviet Physics Doklady, 10*(8), 707–710.

López-Illescas, C., de Moya-Anegón, F., & Moed, H. F. (2011). A ranking of universities should account for differences in their disciplinary specialization. *Scientometrics, 88*(2), 563–574.

Geoff Maslen (2019, August 24), New performance-based funding system for universities, https://www.universityworldnews.com/post.php?story=20190822085127986, accessed on 10th Jan. 2021.

Pislyakov, V., & Shukshina, E. (2014). Measuring excellence in Russia: Highly cited papers, leading institutions, patterns of national and international collaboration. *Journal of the Association for Information Science and Technology, 65*(11), 2321–2330.

Shanghai Ranking's Academic Ranking of World Universities Methodology 2021, https://www.shanghairanking.com/methodology/arwu/2021, accessed on 10th Jan 2021.

Sivertsen, G. (2016). Publication-based funding: The Norwegian model. In Research assessment in the humanities (pp. 79–90). Springer, Cham.

Sörlin, S. (2007). Funding diversity: Performance-based funding regimes as drivers of differentiation in higher education systems. *Higher Education Policy, 20*(4), 413–440.

Torres-Salinas, D., Moreno-Torres, J. G., Delgado-Lopez-Cozar, E., & Herrera, F. (2011). A methodology for institution-field ranking based on a bi-dimensional analysis. *Scientometrics, 88*(3), 771–786.

Uddin, A., Bhoosreddy, J., Tiwari, M., & Singh, V. K. (2016). A Sciento-text framework to characterize research strength of institutions at fine-grained thematic area level. *Scientometrics, 106*(3), 1135–1150.

Zhang, J., Yu, Q., Zheng, F., Long, C., Lu, Z., & Duan, Z. (2016). 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), 967–972.