

A Sciento-text framework to characterize research strength of institutions at fine-grained thematic area level

Ashraf Uddin¹ · Jaideep Bhoosreddy² · Marisha Tiwari³ · Vivek Kumar Singh⁴

Received: 8 August 2015 / Published online: 21 January 2016
© Akadémiai Kiadó, Budapest, Hungary 2016

Abstract This paper presents a Sciento-text framework to characterize and assess research performance of leading world institutions in fine-grained thematic areas. While most of the popular university research rankings rank universities either on their overall research performance or on a particular subject, we have tried to devise a system to identify strong research centres at a more fine-grained level of research themes of a subject. Computer science (CS) research output of more than 400 universities in the world is taken as the case in point to demonstrate the working of the framework. The Sciento-text framework comprises of standard scientometric and text analytics components. First of all every research paper in the data is classified into different thematic areas in a systematic manner and then standard scientometric methodology is used to identify and assess research strengths of different institutions in a particular research theme (say Artificial Intelligence for CS domain). The performance of framework components is evaluated and the complete system is deployed on the Web at url: www.universityselectplus.com. The framework is extendable to other subject domains with little modification.

Keywords Computer science research · Research competitiveness · Field-based ranking · Scientometrics · UniversitySelectPlus

✉ Vivek Kumar Singh
vivekks12@gmail.com

¹ Department of Computer Science, South Asian University, New Delhi, India

² Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA

³ DST-CIMS, Banaras Hindu University, Varanasi, India

⁴ Department of Computer Science, Banaras Hindu University, Varanasi 221005, India

Introduction

There are several university ranking and research performance assessment systems at use at present to assess and rank research/academic performance of institutions across the world. While some ranking systems provide subject wise rankings, none of the well-known ranking systems characterize and rank institutions at fine-grained levels of research themes. For example, Academic Rank of World Universities (ARWU) provides research ranks of institutions for Computer science (CS) subject but it does not distinguish CS research output at finer levels (such as into thematic areas of Artificial Intelligence, Databases, Theory, Algorithms etc.). It is beyond doubt that such a fine-grained assessment of research strength into fine-grained sub-field level of a subject would be more informative and useful. We have, therefore, tried to develop a Sciento-text framework for research performance assessment at fine-grained research area level.

Computer science (CS) domain research output data is used to demonstrate the working of the framework primarily due to two reasons: (a) our affinity with the subject, and (b) the fact that CS has been changing a lot globally because of new innovation and development in technology and computation. It is well known that the institutions do not perform equal in research in all branches of CS as they are not equally equipped with infrastructure and manpower in all areas. It is, therefore, important to characterize research performance in different thematic areas for a better understanding of the research landscape of the discipline and/or an institution. The Web of Science (WoS)¹ data for publication records in CS during 1999–2013 period have been used for the analytical characterization.

The framework design involved two major challenges. First, the WoS data does neither contain full text of articles nor it maps the records into fine-grained thematic areas. It only categorizes the CS research output into seven sub-disciplines based mainly on the source of publications (the genre of journals). The broad area categorization of WoS does not help the purpose since CS is now a vast area of research having evolved over time. For example, ‘artificial intelligence’ sub-field, over the time, has diverged into many thematic branches like data mining, information retrieval, natural language processing etc. We have tried to solve this problem by designing a text classifier to classify each research article into identified finer thematic area(s). Thereafter, we have computed normalized performance indicators for each of the institutions based on their research performance in different thematic areas.

We have implemented the framework and deployed it on the Web scale as a full scale Web-based system, drawing inspiration for broader dissemination from Bornmann et al.’s “Mapping Research Excellence Project” available at www.excellencemapping.net. Our framework identifies strengths as well as time trends of research in different thematic areas of CS at different institutions worldwide. This work is perhaps first work of its kind on a fine-grained level of research performance assessment. The rest of the paper is organized as follows: Sect. 2 presents a brief literature review describing some previous related research work. Section 3 describes the data collection. Section 4 describes in detail the Sciento-text framework designed to do a thematic area-wise analysis. Section 5 presents the results and evaluation followed by a short description of applications in Sect. 6. The paper concludes in Sect. 7 with a summary of the work and its usefulness.

¹ <http://apps.webofknowledge.com>.

Related work and implications

Our work involves two main components: first each research paper is classified into one or more of the thematic areas and secondly relevant indicators for performance of different institutions in different thematic areas of research are computed. The first part required formulation of a suitable classification system. We surveyed the relevant literature to find existing research work on subject/fine-grained thematic area classification task. Some of the previous works have performed individual level classification (Waltman and Eck 2012) and journal level classification (Alwahaishi et al. 2011; Zhang et al. 2010; Janssens et al. 2009; Rafols and Leydesdorff 2009). Waltman and Eck (2012) proposed a three step methodology for constructing a publication-level classification system of science. This involves determining the relatedness of publications based on direct citation relations between publications, clustering publications into research areas and labelling research areas based on the information in titles and abstracts. The study by Alwahaishi et al. (2011) did an analysis of the DBLP publication classification using concept lattices where a journal is represented as a list of topics and the topics are the disciplines being covered by all journals, based on the extracted data from their aims and scopes. Zhang et al. (2010) implemented subject clustering analysis based on ISI category classification. The study focused on the analysis of the information flow among the ISI subject categories and aimed at finding an appropriate field structure of the WoS using the subject clustering algorithm. Unlike the other studies Janssens et al. (2009) used a hybrid text/citation-based method to cluster journals covered by the WoS database. Rafols and Leydesdorff (2009) proposed a content-based and algorithmic classification of journals. In another somewhat related work, Golub (2006) used an automated subject classification of textual Web pages, based on a controlled vocabulary, something conceptually similar to our proposal. Most of this previous work however did not deal with a fine-grained thematic area classification, something which we need for our thematic area classification task. Few other previous works (Singh et al. 2015; Gupta et al. 2011) tried a fine-grained classification of CS domain research output in different thematic areas. The thematic areas in these works, however, are not well distinguished and exhaustive.

To identify thematic areas in a discipline is a difficult task and requires proper understanding of the domain as well as a system to map research papers in different thematic areas. In order to select the target thematic area classes, we used the taxonomy (for CS) provided by Microsoft Academic Search (MAS),² described in detail in the next section. The main reason for selecting MAS is that it is a well distinguished and exhaustive classification system. Further, it uses WoS data for keyword extraction and therefore aligns well with our data collection. We used MAS keyword data to implement an automated classification system for classifying each research paper (publication record) into one or more target thematic classes.

The second component of the framework involved computing performance indicators to identify research strength of institutions in different thematic classes. Some institutions specializing in a given area/subject may have an advantage in the ranking positions when compared with others (Bormmann et al. 2013c). Several approaches are described in the existing literature on research performance assessment and ranking of institutions (such as Liu and Liu 2005; Molinari and Molinari 2008; Lazaridis 2009; García et al. 2012, Uddin and Singh 2015). Different ranking systems use different scientometric indicators and

² <http://academic.research.microsoft.com/>.

different ways to combine them to produce a single composite rank value. Some well-known ranking systems, such as ARWU,³ THE,⁴ QS,⁵ Leiden⁶ etc. also use different scientometric indicators (along with some other parameters) to produce a single value composite rank. Aggregation of different performance indicators into a single value composite rank often involves normalization and/or percentile score computations (Bornmann et al. 2013a, b; Bornmann and Marx 2014; Waltman and Schreiber 2013). There is, however, a huge amount of debate as to which aggregation system works best. There is not one particular accepted aggregation system which is without criticism and drawbacks. Moreover, composite indicators are becoming problematic and also compromise simplicity (Glänzel and Moed 2013). Since, our main focus is on performance assessment of institutions in different thematic areas; we have restricted ourselves to computation of various research performance indicators. Thus, we do a performance assessment of institutions in different thematic areas based on some selected scientometric indicators instead of computing a single composite rank value.

The selection of performance indicators to compute is also not a straightforward task and requires addressing several concerns. We pursued a number of past research works to deliberate upon the issue of selection of performance indicators to compute in our framework. While measuring institutions' performance, the indicators can be aligned to some broader categories such as productivity, citations and collaboration based indicators. One of such paradigms is presented in (Rehn et al. 2007) listing various scientometric indicators. Using total output is a straight forward and simple indicator but cannot offer the aspect of quality at all. But the indicator when restricted to count only high quality journal publications gives some aspect of quality. In scientometric studies, citation based indicators have been seen very popular among researchers. The mean value of class 10 % papers across all institutions is one of the most used estimates for performance indicators like highly cited papers (Bornmann et al. 2014). Use of non-parametric statistics (such as the top-1 %, top-10 %, etc.) is found preferable to central tendency statistics in the case of skewed distribution (Leydesdorff and Bornmann 2012). Experiential distributions, such as the citation distribution have been found to provide better insight into the statistical properties of bibliometric processes than a mean or relative frequency. One such study (Leydesdorff et al. 2011) uses set of six rank percentages defined over certain ranges. Here, each document is assigned a score according to the rank percentage. The proposed R(6) indicator is then defined as the average score over the papers by the unit of analysis. This approach was revised as Integrated Impact Factor (I3) in a subsequent study (Leydesdorff and Bornmann 2011) by summing up the rank scores obtained from the rank percentages of the reference distribution. It has also been proposed that the h index depends on the duration of period being considered so it should be used for same time window and since the h index values are dependent on subject category, they should be used within same discipline (Bornmann and Marx 2011). Article with multiple authors get higher impact (Avkiran and Alpert 2015) and the effect of articles with intellectual collaboration is positive (Ductor 2015). The number of co-authors has also been found to have a positive relationship with the g-index (Bordons et al. 2015). International collaboration can be seen as an important 'amplifier' of measured impact (Van Raan 1998) and also tend to have

³ <http://www.shanghairanking.com/>.

⁴ <https://www.timeshighereducation.co.uk/>.

⁵ <http://www.topuniversities.com/>.

⁶ <http://www.leidenranking.com/>.

higher impact factors as compared to purely native papers (Basu and Aggarwal 2001). Several bibliometric indicators are listed and defined with their uses and pros and cons in a study (Rehn et al. 2007). Here, the indicators are grouped into five different categories. We pursued the studies described above for selection of appropriate indicators, for use in our framework, as described in Sect. 4.2.

Data collection

Bibliometric data

In order to formalize the idea of thematic area based mapping of research strengths, we have chosen CS domain and collected publication data from WoS. The data was downloaded institution-wise for top 530 institutions/research organizations worldwide in CS published during the period 1999–2013. The data collection includes all institutions in the world that have minimum 500 publications during the 15 years period from 1999 to 2013. Some of the institution names in the data are found to be aggregated as a system of institutions (e.g. University of California System) so we have dropped those names. Finally we were left with 444 institutions with total 498,488 publication records (as in August 2014). Each record has 60 meta-data fields such as authors, title, affiliation, abstract, keywords, publication source, and reference list.

Classified control-term dictionary

For the thematic area classification, we have collected the MAS data. MAS is an academic publication search engine provided by Microsoft Research. It indexes research publications from journals and conferences in various fields of research and provides features like data mining, entity linking and visualisation. MAS provides API-level access to their data through their Microsoft Azure Cloud Computing service. For CS domain, MAS classifies all the publications into 24 academic specialisations or thematic classes. MAS also provide a comprehensive list of high frequency keywords corresponding to each of the 24 thematic classes. We have used the keyword data mappings to different thematic classes to classify the publication records in the downloaded dataset into different thematic areas/classes. A total of 37,208 keywords were collected with their occurrence frequencies in different thematic classes. The occurrence frequency data of MAS was retrieved on 25 April 2015. The author keywords and other selected fields of a paper are then matched with this keyword list in a systematic sequential manner for the thematic area classification task.

Institution metadata

We have also collected and organized data and information about institutions in the dataset from various sources. This data is displayed in the web interface for ready reference and accessibility purposes. We have collected Wikipedia infobox data for all the institutions where it was available. In addition, we have collected the geographical latitude and longitude coordinates of each institution. We have also collated institution phone numbers, email addresses, postal addresses and other such information for the benefit of the end users. Collectively, this meta-data information is useful for Web application functionality.

The Sciento-text framework

We have designed a computational framework, called Sciento-text framework, that combines scientometric and text analytics based techniques. The thematic area classification part is addressed by the text analytics component and the performance assessment part by using scientometric methodology. We describe below the two main parts of the framework.

Thematic area classification

In order to classify a research paper to belong to a particular thematic area, a mapping between research papers to thematic areas is required. Each research paper publication record) is evaluated to belong to one or more of the thematic areas. Author keyword, title and abstract field data of research papers is extracted and used for classification task. A four step algorithm is implemented for the purpose and is defined as follows: Let,

$D = \text{set of documents (research papers)},$

$C = \text{set of classes (thematic areas)}$

$K = \text{set of keywords from MAS.}$

We perform the classification through application of following steps:

Step 1 Keyword frequency matrix

First of all, a $|K| \times |C|$ keyword frequency matrix F is constructed.

Step 2 Normalization of frequency

The keyword distribution of different thematic areas is not uniform so a way to find relative importance of each keyword with respect to the thematic classes needs to be defined. For this purpose, we have normalized the frequency values (by dividing the keyword frequency values by the maximum number of occurrence of that keyword in any thematic area of publications). If a keyword is very common in many thematic research areas, it does not carry much specificity for those areas. To reduce the effect of such commonality, Inverse Document Frequency (IDF) value of keywords is multiplied with corresponding frequency values. More precisely, following mathematical expressions are computed:

For each $k \in K$,

$$IDF_k = \log\left(\frac{|C|}{D_k}\right)$$

$$m_k = \max\{F[k, j], j = 1 \text{ to } |C|\}$$

For each class $c \in C$,

$$M[k, c] = EDF_k * \left(\frac{F[k, c]}{m_k}\right)$$

Step 3 Class scores

For each research paper (document), the keywords are extracted from title, author keywords and abstract. The extracted keywords are denoted as DT_d, DK_d, DA_d , respectively. Three scores (title, author keywords and abstract) for each set are calculated and combined using the following formula:

For each document $d \in D$,

$$DT_d, DK_d, DA_d \subseteq K$$

For each class $c \in C$,

$$Score(DT_d) = \frac{1}{|DT_d|} * \sum_{K_i \in DK_d} M[i, c]$$

$$Score(DK_d) = \frac{1}{|DK_d|} * \sum_{K_i \in DT_d} M[i, c]$$

$$Score(DA_d) = \frac{1}{|DA_d|} * \sum_{K_i \in DA_d} M[i, c]$$

$$S[d, c] = \frac{4}{7} * Score(DT_d) + \frac{2}{7} * Score(DK_d) + \frac{1}{7} * Score(DA_d)$$

We see that different weights are given to different scores, which are then combined as a linear weighted sum into a single score. The weightage given to occurrence in title is more as compared to occurrence in author keywords and/or abstract. These weights are distributed in the ratio of 4:2:1. This is done by making an intuitive assumption that the keywords used in the title are most decisive for class membership followed by those in the author keyword list followed by abstract. These calculations thus lead to $|D| \times |C|$ score matrix(S). This matrix contains membership score of each document for each of the 24 thematic classes.

Step 4 Class assignment

This is the final step of the classification algorithm. Till now each document is assigned a vector of scores (for all the thematic classes). However, since we want crisp assignment of each document into thematic classes, we use a threshold value to decide which thematic class(es) a document may be assigned to. A document is normally assigned a class for which the maximum score is obtained. But there may be other classes having scores close to the maximum score. To resolve the issue a threshold value is chosen. Now a document for which other class scores are close enough to the maximum score may be assigned to more than one class. More precisely, the threshold and classes are determined in the following manner:

For each document d ,

$$class_d = \{ \}$$

$$\theta_d = \max\{S[d, j], \quad j = 1 \text{ to } |C|\} - sd\{ S[d, j], j = 1 \text{ to } |C|\}$$

For each class $c \in C$,

$$\text{If } S[d, c] \geq \theta_d, \text{ add } c \text{ to } class_d$$

Here, sd refers to standard deviation. We observe that the threshold value is computed by reducing the maximum value by the standard deviation of the all class values for the document. In this way, we classify each document (publication record) into one or more thematic classes. For the CS data that we have used, 72.84 % research papers are assigned to a single class whereas 18.71 % are classified into two classes. About 8.45 % of research papers are classified into three or more than three classes.

Selecting and measuring performance indicators

After completing the thematic area classification, it is required to identify research strengths (in terms of scientometric indicators) for each institution in each thematic area. Based on the studies mentioned in the ‘literature review’ section, we have selected a total of ten indicators and grouped them into related categories, as described below. Some of these indicators are also being used by the well-known ranking schemes for the purpose of assessment of research performance. For example, total number of publications in journals indexed in Thomson Reuters’ SCIE and SSCI is used in ARWU ranking. The top 10 % most frequently cited papers are used as research performance indicators in Leiden and SCImago ranking schemes. Similarly, both the total citations and average citations are used in Leiden ranking, while a normalized variation of it is used in THE, QS and Leiden ranking schemes. Total number of publications in top journals is used by both ARWU and SCImago, though they differ in the methods of defining the top journals. International collaborative publications are used in Leiden, THE and SCImago ranking schemes. All three of them use the ratio of international collaborative publications to the total number of publications, thus making it a size-independent indicator while ours is a size-dependent one. The list of indicators used in the present study along with their categorisation is as follows:

Publication based indicators

TP: total number of publications; **TP_{top_journals}**: number of publications in first quarter journals according to the impact factors published by WoS Journal Citation Report 2013.

Citation based indicators

TC: total number of citations, defined as $TC = \sum_{i=1}^{TP} \text{Citation}_i$; **CPP**: citations per publications, defined as $CPP = TC/TP$; **HiCP**: number of publication in top 10 % cited set, where top 10 % cited set is obtained from reference publication set. In this study, the whole publication set is considered as reference publications; **Cited %**: it is defined as the percentage of cited publications to total publications; **I3**: it is sum of weighted frequency based percentile ranks as defined in (Leydesdorff and Bornmann 2011). The citations are weighted in accordance with the percentile rank class of each publication in an integrated impact indicator, $I3 = \sum_i x_i * f(x_i)$. In this formula, x represents the percentile value and $f(x)$ the frequency of this rank; **h index**: a scientist has index h if h of his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have $\leq h$ citations each (Hirsch 2005).

Collaboration indicators

Co-authorship: it is defined as the average number of authors per publication; **ICP**: number in internationally collaborated paper. A publication is considered as an ICP instance if there are authors affiliated to more than one country.

Determining field normalized research strengths

Instead of showing absolute values for a thematic area in an institution, it is more meaningful to see the comparative strength (using field normalization) in the given thematic area. This is due to the fact that for an institution, absolute value of an indicator does not give proper insight about the relative strength when compared with other disciplines in the same institution. The variation in productivity levels for different thematic areas is responsible for this. Thus, it is required to normalize an indicator value considering the average productivity of the thematic area as base value. For a particular indicator I and thematic area t , the field normalized value of research performance of an institution i can be defined as:

$$NF(I_{t,i}) = n * \frac{I_{t,i}}{\sum_{j=1}^n I_{t,j}}, \quad \text{where } n \text{ is the number of institutions}$$

The field normalized strength is computed for each of these indicators to show relative strength of the institution in different thematic areas. The base value, against which the indicator value is compared, is the average indicator value.

The indicator values for each institution are normalized along all thematic areas. For a particular indicator I and thematic area t the normalized score of an institution i can be defined as:

$$N(I_{t,i}) = \frac{I_{t,i}}{\max_{j=1 \text{ to } n} I_{t,j}}, \quad \text{where } n \text{ is the number of institutions}$$

The institutions are ranked for a particular thematic area on each of the individual indicators computed.

Results and evaluation

We now show the computational results and outcomes of the Sciento-text framework on the CS domain research publication data for leading institutions worldwide. First of all we report classification accuracy. Then we show example results for field-normalized research strengths on different scientometric indicators.

Accuracy of classification

Out of the 498,488 documents, our classification system was able to classify a total of 404,167 documents (~ 81 %) into one or more thematic classes. We tried to compute the accuracy of classification. For this purpose, we have taken a randomly selected subset of classified documents (total 480 classified documents, 20 documents for each thematic area). These documents were then annotated by an independent annotator. Out of these 480 randomly selected documents, a total of 441 documents are found to be correctly classified indicating the classification accuracy to be ~92 %, which is a reasonably good accuracy level for the classification.

Field normalized research strengths

We have identified the research strengths of different institutions in all the 24 thematic areas of CS research. First, we show the distribution of research output of all the institutions taken together into 24 thematic areas in Fig. 1. We observe that ‘Artificial Intelligence’ leads the field of CS significantly followed by ‘Algorithms & Theory’ and ‘Networks & Communications’. Secondly, we show the normalized research strengths of three prominent institutions in CS research in 24 different thematic areas. The Fig. 2 shows field normalized research strength of three institutions. The figure clearly shows that Carnegie Mellon University (CMU) performs better than other two in ‘Operating Systems’, ‘Programming Languages’, ‘Computer Education’ and ‘Software Engineering’. However, a look at several such charts shows that research strengths of institutions vary in different thematic areas. The Table 1 presents the screenshot of performance assessment results in ‘Artificial Intelligence’ and the Table 2 presents the screenshot of performance assessment in ‘Software Engineering’.

Application

We have implemented and deployed our Sciento-text framework as a Web-based system, which is available on the website: www.universityselectplus.com. The system incorporates the functionalities of classification, computation of field normalized research strengths and comparison of different institutions across different thematic areas. At present the system has analytical results computed for CS domain only but we are working towards generalizing it by taking publication data for other disciplines as well. The purpose of the web-based application framework is to make the inferences derived from our research easily

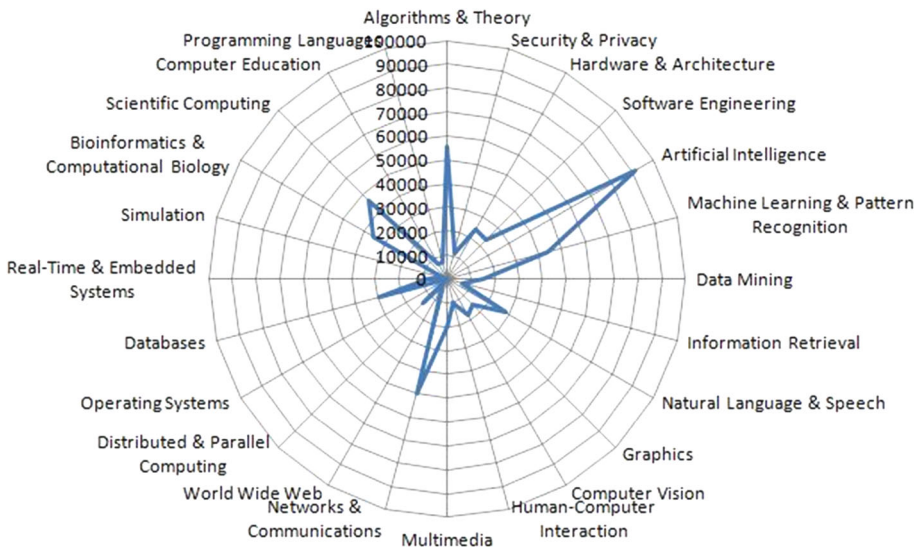


Fig. 1 Thematic area characterization of CS research output data (1999–2013)

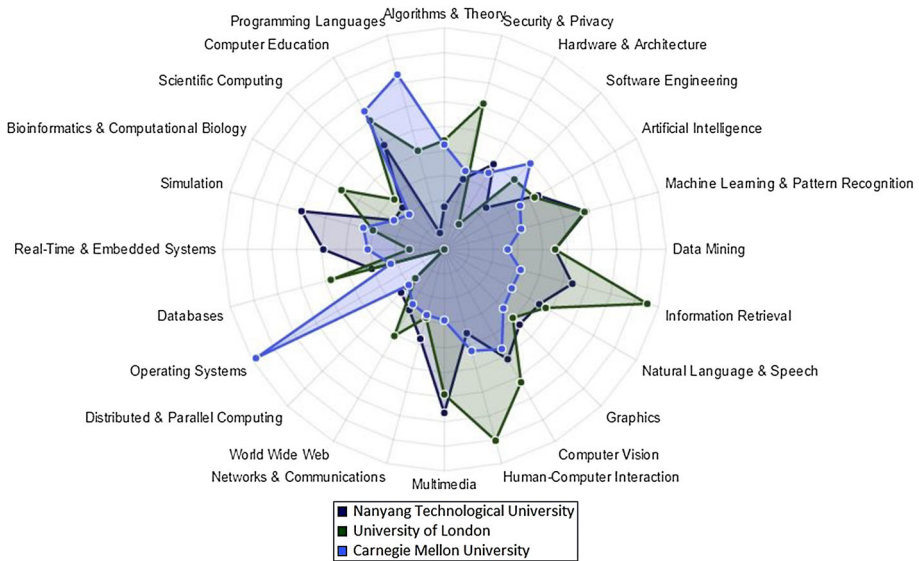


Fig. 2 Field normalised research output strength (TP) of three selected institutions

accessible to prospective students and researchers and to serve as an informative and helpful tool for the academic and research community.

The primary target user groups, for which we developed the Web-based system, are prospective and current undergraduate and graduate students, researchers, policy makers and administrators. From a student’s point of view, the application gives in-depth analysis of the research being done at all the leading academic institutions, including identification of research strengths in different thematic areas. It is also possible to compare research performance of different institutions in one or more thematic areas. The application also provides temporal trends of research in different thematic areas and/or different institutions. The results produced also provide an account of most productive and cited authors in a thematic area and/or an institution. This information may also be useful for policy makers and administrators who can use the system to know the status of research in different institutions and thematic areas and consequently take decisions about research policy formulation and funding schemes.

The web application of the system is built on open-source technologies like PHP and HTML5 + CSS3 and Javascript. In the system many forms of results can be visualized. The Fig. 3 shows the home page of the web implementation. The web-based system provides different results, such as comparison of research performance assessment of institutions (worldwide or in a region/country) in one or more thematic areas, most productive and most cited authors in a thematic area in an institution etc.

Table 1 Research performance of 10 most productive institutions in Artificial Intelligence thematic area (1999–2013)

Institution	Performance indicators									
	Publication based			Citation based			Collaboration based			ICP %
	TP	TP _{top_journals}	TC	CPP	HiCP	Cited %	I3	h Index	Co-authorship	
1 Nanyang Technological University	904	351	16,247	17.97	13.5	85.51	1971	58	4.13	43.14
2 University of London	870	294	13,198	15.17	12.18	82.18	1750	54	4.39	44.48
3 Tsing Hua University	843	312	10,826	12.84	9.96	78.77	1634	47	4.31	24.79
4 National University of Singapore	769	295	12,215	15.88	13.65	85.31	1647	55	4.19	42.26
5 Carnegie Mellon University	730	290	16,343	22.39	17.53	86.44	1686	60	4.33	33.42
6 City University of Hong Kong	717	357	14,256	19.88	18.83	87.73	1677	60	4.24	27.48
7 Hong Kong Polytechnic University	676	341	9106	13.47	11.83	87.72	1421	42	4.31	26.92
8 Shanghai Jiao Tong University	639	213	5963	9.33	7.67	75.74	1127	38	4.37	20.81
9 Massachusetts Institute of Technology	629	244	13,666	21.73	15.58	83.94	1401	55	4.36	34.5
10 University of Granada	596	259	12,215	20.49	16.78	87.25	1320	53	4.68	21.31

Table 2 Research performance of 10 most productive institutions in Software Engineering thematic area (1999–2013)

Institute	Performance indicators										
	Publication based			Citation based			Collaboration based				
	TP	TP _{top_journals}	TC	CPP	HiCP	Cited %	I3	h	Index	Co-authorship	ICP %
1 Carnegie Mellon University	259	48	4637	17.9	19.69	83.4	613	33		4.36	35.14
2 University of London	211	26	2339	11.09	16.11	80.57	449	27		4.43	44.55
3 University of Illinois At Urbana-Champaign	184	26	2202	11.97	14.13	80.98	400	23		4.57	34.24
4 Eindhoven University of Technology	178	20	1847	10.38	8.43	77.53	342	20		4.03	41.57
5 National University of Singapore	168	37	1377	8.2	9.52	74.4	321	21		4.35	54.17
6 University of Texas At Austin	164	26	1425	8.69	11.59	77.44	327	21		4.24	19.51
7 Imperial College London	161	32	1867	11.6	11.8	81.37	329	21		4.37	45.96
8 University of New South Wales	153	26	1506	9.84	9.8	84.31	311	19		4.2	97.39
9 Massachusetts Institute of Technology	151	26	1923	12.74	13.91	84.77	329	22		4.34	35.76
10 Technical University of Munich	150	18	1192	7.95	7.33	79.33	277	18		4.19	38

University Select Plus

Domain: Choose a domain

Specialisation: No selection

Geographic Scope: Global

Choose Year Range: Year 1999 Year 2013

Submit

UNIVERSITIES 444

SUBJECTS 1

SPECIALISATIONS 24

Searching for higher education?

Look no further - University Select Plus: Your one-stop gateway to education.

University Select Plus is a scientometric tool developed by Computer Science researchers to help guide prospective students choose the right academic path.

1. Select your subject, specialisation and country you want to study in from the sidebar.
2. Compare up to three institutes on various parameters of research and academic performance.
3. Get deep insights about your specific institute like never before.

It's as easy as 1,2,3.

Fig. 3 The screenshot of the Web-based system homepage

Conclusion

We have designed a Sciento-text framework for a fine-grained analysis and research performance assessment of institutions in different thematic areas. The working and suitability of the framework is demonstrated with CS discipline as a case by mapping and analysing the CS research output into different thematic areas. Our framework provides a research theme specific assessment of research competitiveness of institutions. None of the renowned ranking systems provide such a deeper insight into a particular discipline, though some of them do compute ranks at a broader subject level (such as CS as a whole). Our system uses a generalized framework which can perform thematic area level analysis for any discipline of research if the publication data and a classification system are available. The institutions can be ranked on different scientometric indicators derived from their research performance in different thematic areas. The entire analytical framework is evaluated and deployed on the Internet at <http://www.universityselectplus.com>. All the 444 institutions are ranked in all 24 thematic areas of CS. Country level and region level analysis results are available. The results produced are quite comprehensive, informative and useful.

Our application is expected to be helpful for prospective students in making better decisions regarding their choice of research discipline and institution. The researchers and policy makers will also benefit from the system in a number of ways. The funding agencies can take a call to fund more to high performing institutions in a particular area and create a centre for excellence in a particular area. However, our framework does not compute a single value composite rank of research competitiveness at present. We are working towards providing a single value composite rank as an extension of the present work. Further, the classification accuracy for class assignment, which is around 81 % at present, can be improved a bit. To the best of our knowledge this work is the first of its kind with a

comprehensive framework for fine grained research performance analysis, thematic characterization and thematic area based assessment of research competitiveness of leading world institutions. This framework is a general design that can be suitably extended to other disciplines of research.

Acknowledgments This work is supported by research grants from Department of Science and Technology, Government of India (Grant: INT/MEXICO/P-13/2012) and University Grants Commission of India (Grant: F. No. 41-624/2012(SR)). A preliminary version of this work was presented in 20th Science Technology Indicators Conference in Sep. 2015 at Lugano, Switzerland.

References

- Alwahaishi, S., Martinovič, J., & Šnášel, V. (2011). Analysis of the DBLP Publication Classification Using Concept Lattices. *Digital enterprise and information systems* (pp. 99–108). Berlin: Springer.
- Avkiran, N. K., & Alpert, K. (2015). The influence of co-authorship on article impact in OR/MS/OM and the exchange of knowledge with Finance in the twenty-first century. *Annals of Operations Research*, 235(1), 1–23.
- Basu, A., & Aggarwal, R. (2001). International collaboration in science in India and its impact on institutional performance. *Scientometrics*, 52(3), 379–394.
- Bordons, M., Aparicio, J., González-Albo, B., & Díaz-Faes, A. A. (2015). The relationship between the research performance of scientists and their position in co-authorship networks in three fields. *Journal of Informetrics*, 9(1), 135–144.
- Bornmann, L., Leydesdorff, L., & Mutz, R. (2013a). The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. *Journal of Informetrics*, 7(1), 158–165.
- Bornmann, L., Leydesdorff, L., & Wang, J. (2013b). Which percentile-based approach should be preferred for calculating normalized citation impact values? An empirical comparison of five approaches including a newly developed citation-rank approach (P100). *Journal of Informetrics*, 7(4), 933–944.
- Bornmann, L., & Marx, W. (2011). The h index as a research performance indicator. *EurSci Ed*, 37(3), 77–80.
- Bornmann, L., & Marx, W. (2014). How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations. *Scientometrics*, 98(1), 487–509.
- Bornmann, L., Moya Anegón, F., & Mutz, R. (2013c). Do universities or research institutions with a specific subject profile have an advantage or a disadvantage in institutional rankings? *Journal of the American Society for Information Science and Technology*, 64(11), 2310–2316.
- Bornmann, L., Stefaner, M., de Moya Anegón, F., & Mutz, R. (2014). Ranking and mapping of universities and research-focused institutions worldwide based on highly-cited papers: A visualisation of results from multi-level models. *Online Information Review*, 38(1), 43–58.
- Ductor, L. (2015). Does co-authorship lead to higher academic productivity? *Oxford Bulletin of Economics and Statistics*, 77(3), 385–407.
- García, J. A., Rodríguez-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.
- Glänzel, W., & Moed, H. F. (2013). Opinion paper: Thoughts and facts on bibliometric indicators. *Scientometrics*, 96(1), 381–394.
- Golub, K. (2006). Automated subject classification of textual Web pages, based on a controlled vocabulary: Challenges and recommendations. *New Review of Hypermedia and Multimedia*, 12(1), 11–27.
- Gupta, B. M., Kshitij, A., & Verma, C. (2011). Mapping of Indian computer science research output, 1999–2008. *Scientometrics*, 86(2), 261–283.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569–16572.
- Janssens, F., Zhang, L., De Moor, B., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing and Management*, 45(6), 683–702.
- Lazaridis, T. (2009). Ranking university departments using the mean h-index. *Scientometrics*, 82(2), 211–216.

- Leydesdorff, L., & Bornmann, L. (2011). Integrated impact indicators compared with impact factors: An alternative research design with policy implications. *Journal of the American Society for Information Science and Technology*, 62(11), 2133–2146.
- Leydesdorff, L., & Bornmann, L. (2012). The integrated impact indicator (I3), the top-10% excellence indicator, and the use of non-parametric statistics. *Research Trends*, 29, 5–8.
- Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the tables on citation analysis one more time: Principles for comparing sets of documents. *Journal of the American Society for Information Science and Technology*, 62(7), 1370–1381.
- Liu, N. C., & Liu, L. (2005). University rankings in China. *Higher Education in Europe*, 30(2), 217–227.
- Molinari, A., & Molinari, J. F. (2008). Mathematical aspects of a new criterion for ranking scientific institutions based on the h-index. *Scientometrics*, 75(2), 339–356.
- Rafols, I., & Leydesdorff, L. (2009). Content-based and algorithmic classifications of journals: Perspectives on the dynamics of scientific communication and indexer effects. *Journal of the American Society for Information Science and Technology*, 60(9), 1823–1835.
- Rehn, C., Kronman, U., & Wadskog, D. (2007). Bibliometric indicators—definitions and usage at Karolinska Institutet. *Karolinska Institutet*, 13, 2012.
- Singh, V. K., Uddin, A., & Pinto, D. (2015). Computer science research: The top 100 institutions in India and in the world. *Scientometrics*, 104(2), 539–563.
- Uddin, A., & Singh, V. K. (2015). A quantity–quality composite ranking of Indian institutions in computer science research. *IETE Technical Review*, 32(4), 273–283.
- Van Raan, A. (1998). The influence of international collaboration on the impact of research results: Some simple mathematical considerations concerning the role of self-citations. *Scientometrics*, 42(3), 423–428.
- Waltman, L., & Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12), 2378–2392.
- Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology*, 64(2), 372–379.
- Zhang, L., Liu, X., Janssens, F., Liang, L., & Glänzel, W. (2010). Subject clustering analysis based on ISI category classification. *Journal of Informetrics*, 4(2), 185–193.