A Sciento-Text Framework for Fine-grained Characterization of the Leading World Institutions in Computer Science Research

Ashraf Uddin[1], Sumit Kumar Banshal[2], Khushboo Singhal[3] and Vivek Kumar Singh[4]

[1]*mdaakib18@gmail.com,* [2]*sumitbanshal06@gmail.com,* [3]*khushbusinghal18@gmail.com,* [4]*vivek@cs.sau.ac.in*
Department of Computer Science, South Asian University, New Delhi (India)

Introduction

This paper describes our experimental framework for a text analysis based fine-grained characterization of leading world institutions in Computer Science (CS) research. Though the present paper uses CS research output data from Web of Science, it can be extended and applied to any discipline and data source. The existing well-known ranking systems, such as ARWU[1],Times Higher Education World University rankings[2], QS World University Rankings[3], SIR[4], Leiden Ranking[5] and Webometrics[6], only present an overall (or for a whole discipline) rank of institutions. These rankings may not be helpful if one is interested in knowing centers of excellence in research in a particular area (say Artificial Intelligence or Software Engineering in CS). Such fine-grained characterization could be very useful for different purposes. Prospective students looking to work in a particular specialized area may look at the fine-grained characterization and select institutions accordingly. Academicians or industry professionals looking for collaboration in a particular area can use the information for selecting potential institutions for collaboration. Similarly, funding agencies and policy making bodies in a country may identify institutions strong in different specialized areas of research. The other advantage of this kind of sciento-text characterization is that it is completely automated, verifiable and does not use any perceptual scores for ranking (such as reputation survey and perceptual scores of QS). Our system thus proposes a framework that uses scientometric data to produce a fine-grained research strength characterization of institutions and to rank them in order of their research excellence in a particular area.

Data Collection

We have demonstrated the working and suitability of our approach for CS domain. We obtained research output data for CS domain for the period 1999 to 2013 indexed in Web of Science (WoS). The data has been collected through an institution-wise search and we collected data for top 100 most productive institutions. A total of 261,154 records were obtained. This data constitutes about 34% of the total worldwide CS domain research output (784,920 records in total) for the period 1999-2013.

Sciento-Text Based Analytical Framework

Since our main objective is to produce a fine-grained characterization and consequential rankings, we had to first assign every research output to one or more particular research specialization. We identified a total of 11 major thematic areas (specializations) in CS domain research output. The 11-classes are based on perusal of data, some recent work (Gupta et al., 2011; Uddin et al., 2015) and recent research trends in the discipline. We processed each record in the data, extracted its 'title', 'author keywords' and 'abstract' fields and obtained the text contents of these fields. For classifying a record (research paper) to belong to one or more of the 11 thematic areas (specializations), a simple Naïve Bayes (NB) text classifier is used. The names of the 11 classes are embedded in table 1. For obtaining training data for the NB classifier, we used a keyword-match strategy for a part of the data. First of all, we created a term-profile for each thematic area (through a manual annotation by three independent annotators). Then, each record is checked for occurrence of any term from the term-profile of the 11 thematic classes, in its 'author keyword', 'title' and 'abstract' fields, in a sequential manner. Those records which get an exact match of keywords with one or more of the 11 thematic classes are assigned that class label. The assigned records then serve as training set for NB classifier, which is then used to classify the remaining unclassified records. In this manner, we classify each record to belong to one or more of the 11 thematic classes. After assigning thematic class to each record, we partitioned the data into 11 groups. Now, we have research output data for each of the major thematic areas (specializations) from the 100 most productive institutions of the world. This information is now used to first produce a plot of the research output landscape of the 100 most productive institutions and then to identify top ranking institutions in all the thematic areas. For ranking we use a simple average of scientometric indicator values for these

[1] http://www.shanghairanking.com/
[2] http://www.timeshighereducation.co.uk/world-university-rankings/
[3] http://www.topuniversities.com/university-rankings
[4] http://www.scimagoir.com/
[5] http://www.leidenranking.com/
[6] http://www.webometrics.info/

Table 1. Thematic Area Wise Top Ranking Institutions.

AI	CT	CHA	CN	CSA	CG	DBMS	IM	OS	SIP	SE
NTU	NTU	INRIA	INRIA	UCB	INRIA	NTU	TU	INRIA	NTU	INRIA
UCB	MIT	IBM	NTU	INRIA	SJTU	HU	INRIA	TU	UL	UCB
TU	INRIA	TU	UCB	KL	NTU	INRIA	MS	KL	UCB	HU
MS	UL	NTU	TU	NTU	UT	MIT	NUS	HKPU	NUS	UL
UGR	UM	GIT	CUHK	UL	UL	UL	HU	IBM	UIUC	MIT
CUHK	UTA	UCB	HIT	CMU	UW	NUS	NTU	UM	MS	NTU
INRIA	PSU	INTEL	UNC	TU	KL	MS	SU	UW	INRIA	UNC
HKPU	CMU	MS	UL	GIT	TU	MPG	CUHK	UCSD	TAU	UMCP
HU	UCL	PUC	SU	MIT	CUHK	CU	UL	NTU	TU	TU
UL	SU	CMU	GIT	MPG	IBM	IBM	MIT	UCB	KL	IBM

AI : Artificial Intelligence, CT: Computation Theory, CHA: Computer Hardware & Architecture, CN: Computer Networks ,CSA: Computer Software & Applications, CG: Cryptography, DBMS: Database Management System, IM: Internet & Multimedia, OS: Operating System, SIP: Signal & Image Processing, SE: Software Engineering

institutions, namely TP (Total Papers), TC (Total Citations), ACPP (Average Citations Per Paper), and HiCP (Highly Cited Papers). The absolute scores are first normalized to 0-100 range and then a simple arithmetic average is computed. One such similar ranking work (without thematic areas) is presented in a past literature (Ma et al., 2008).

Results and Conclusion

Our framework produces a detailed characterization of research output along the major research themes by the 100 most productive institutions of the world. The Figure 1 presents a plot of TP and TC values along the 11 research themes for the whole set of 100 institutions. Top ranking institutions identified in all 11 thematic research areas for the given period are listed in table 1. It can be seen that many of the institutions are almost available in each list but with different rank positions. Thus the presented results verify the importance of ranking institutions in different thematic areas rather than doing it for a broader research field. The paper thus presents an interesting framework for fine-grained characterization of leading world institutions and to identify the top ranking institutions in different thematic areas of CS domain. The work is extendable to other disciplines and data sources. The work may benefit more if we would have incorporated the number of researchers and graduate students for better insightful result but unfortunately obtaining those data for each institution is cumbersome and time consuming. See http://www.viveksingh.in/publications/issi2015/appendix.pdf for the full names of institutions.

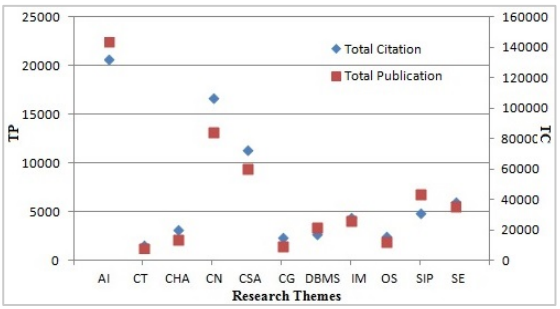

Figure 1. Thematic Area Wise Research Output and Citations.

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)).

References

Gupta, B.M., Kshitij, A., & Verma, C. (2011). Mapping of Indian computer science research output, 1999–2008. *Scientometrics*, *86*(2), 261–283.

Ma, R.., Ni, C., & Qiu, J. (2008). Scientific research competitiveness of world universities in computer science. *Scientometrics*, *76*(2), 245–260.

Singh, V.K., Uddin, A., & Pinto, D. (2015). Computer Science Research: The Top 100 Institutions in India and in the World. Submitted in *Scientometrics*.