Reputation in Computer Science on a per Sub-area Basis

Alberto Ueda
Computer Science Department
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
ueda@dcc.ufmg.br

Marlon Dias
Computer Science Department
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
msdias@dcc.ufmg.br

Berthier Ribeiro-Neto
Computer Science Department
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
berthier@dcc.ufmg.br

Nivio Ziviani
Computer Science Department
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
nivio@dcc.ufmg.br

Sabir Ribas
Computer Science Department
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
sabir@dcc.ufmg.br

ABSTRACT

In this work we study the reputation of venues and research groups in Computer Science (CS) with focus on its sub-areas. We adopt the 37 sub-areas defined by Microsoft Academic as starting point and focus on the CS departments in the USA. We explore answers to two basic questions, as follows. First, how do the reputation of venues and research groups vary per sub-area? Second, how does the reputation of venues and research groups per sub-area compares to their reputation when sub-areas are not taken into account? Both answers are important for university officials, funding agencies directors, and government officials who need to decide how to allocate limited research funds.

For evaluating the reputation of venues and research groups we rely on a reputation framework we introduced in previous work and its derived metric, called P-score (for Publication Score). P-scores rely on a Markov network that models relations among researchers, among researchers and the venues they publish in, and among the venues themselves (through paper citations). By properly setting a configuration parameter, the network allows computing P-scores without consideration to the relations among venues i.e., without using citation information. Further, the P-scores so computed present good correlation with citation based metrics and yield good results, as demonstrated in previous work.

We run several experiments in which we compare the reputation of venues and research groups per sub-area, using P-scores and citation counts. The results suggest that P-scores yield better results. We also study how to combine scores per sub-area on an overall reputation score and show that the results are again superior (to those produced directly without consideration to sub-areas). Most important, analysis of reputation on a per sub-area basis yields additional insights into the reputation of venues and research groups that are useful when deciding how to allocate research funds.

ICTIR'17, Amsterdam, Netherlands

DOI: 10.1145/nnnnnnn.nnnnnnn

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**;

KEYWORDS

Academic search, random walks, reputation flows

ACM Reference format:
Alberto Ueda, Marlon Dias, Berthier Ribeiro-Neto, Nivio Ziviani, and Sabir Ribas. 2017. Reputation in Computer Science on a per Sub-area Basis. In *Proceedings of ACM International Conference on the Theory of Information Retrieval, Amsterdam, Netherlands, October 2017 (ICTIR'17),* 4 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As the number of possibilities to a given entity (e.g. a person, a company, or a institution) grows up, the more important becomes the task of ranking these options and present a reasonable result to the entity. For instance, one wants to find a relevant source of information in the Web about a specific topic of interest, among trillions of documents available. Other examples of use cases could be: a person looking for great places to visit during her vacation; companies searching for the most compatible candidates to the jobs they are offering; or a research institution interested in knowing how is the public perception about itself in terms of reputation.

In a lower level of abstraction, all these tasks are composed by other smaller challenges, including i) to gather all the information found of a specific domain; ii) to structure that data; iii) to propose theoretical models in order to return the most relevant items to the entity, based on its interests and context; and iv) evaluate each one of these models properly.

One of the specific domains where this problem applies is the academic research domain. We can model this context using a few entities: researchers, research groups, papers, and publication venues. Each one of these entities is subject to different ranking methods, according to its type and the criteria one intends to measure. We could be interested in, for instance, retrieving the most popular venues (i.e. publication conferences and journals) in the broad area of Computer Science; or finding the authors in Chemistry whose work is the most related to a specific topic of study (a task generally referred as *expert search*); or deciding which national institution is the most suitable for the allocation of research funds for the next years.

Reputation is one of these criteria commonly used on academic rankings. Although ill-defined, the reputation of a entity reflects the estimation in which this entity is held by the public, developed over time. At first, the entities associated with a high reputation are more indicated to receive new research funds, grants, academic awards, graduate students, and other incentives, than less reputable entities within the same community [9].

Several studies of how to measure the reputation of academic entities were proposed, including citation-based metrics, machine learning techniques, and random walks on the academic graph. Some of them (e.g. H-index and Impact Factor) are being used by national institutes as standard methods for measure academic productivity and research impact in the broad areas of knowledge.

The problem is that most of these metrics does not leverage the specific domain where these entities are. General purpose rankings of national Computer Science departments, for instance, may not reflect important information about these departments from the perspective of sub-areas. For example, if a company wants to invest funds in a reliable research group working on the sub-area of Human Computer Interaction, general rankings of the broad area of Computer Science surely will not be enough to make the decision of which is the better department to invest in.

Other noteworthy scenario is comparing two different researchers, one of them belonging to a sub-area *A* and the other belonging to sub-area *B*. If we assume that it is inherently harder to publish articles in *A* than in *B*, it seems natural that the metrics used to ranking these researchers must be distinct, or, at least, consider those differences between sub-areas into account. Otherwise, the comparison between them would be fatally unfair.

In this work we suggest methods to get insights about the academic entities on a per sub-area basis, focusing in the area of Computer Science. For doing that, we extend a reputation framework to the context of sub-areas. In particular, we aim to answer the following research questions:

Q1. How do the reputation of venues and research groups vary per sub-area?

Q2. How does the reputation of venues and research groups per sub-area compares to their reputation when sub-areas are not taken into account?

In summary, the main contributions of this work are:

(1) An extension of the reputation framework that could shed a light on the reputation of entities by using a context-aware approach, where only entities in the same context are compared;
(2) A validation of that extension on an real-world setting of problem of measuring reputation: ranking academic entities in Computer Science;
(3) A few insights into the current reputation of publication venues and US research groups in Computer Science, from the perspective of sub-areas.

The remainder of this paper is structured as follows. In Section 2 we describe the related work on reputation models and some instantiations of these models in academic search tasks. In Section 3, we present the theoretical concepts supporting our approach, by describing the key ideas of the reputation model we used in this work and its derived metric, called P-score. In that section we also formalize the strategies we adopted to study the academic data on a per sub-area basis. The academic dataset and the experimental methodology we adopt are described in detail in Section 4. The main results we obtained in the academic setting are present in Section 5. In Section 6 we discuss the key contributions of this work and directions for further research.

2 RELATED WORK

Martins et al. [7]

Goncalves et al. [3]

Wainer et al. [11]

Hoonlor et al. [4]

Gollapalli et al. [2]

Gonçalves et al. [1]

Lima et al. [6]

ACM Special Interest Groups [1]

Google Scholar [2]

Microsoft [3]

AMiner [4], maintained by Tang et al. [10]

CiteSeerX [5], directed by Giles et. al. [5]

Semantic Scholar [6], mantained by the Allen Institute for Artificial Intelligence.

CS Rankings [7], maintained by Berger of UMass.

Citation-based Metrics

Academic Search

P-score and Reputation Flows

Ribas et al. [9]

Ribas et al. [8]

3 FRAMEWORK

3.1 P-score

3.2 Venue Ranking

3.3 Group Ranking

Evaluating a group is a task that demands inspection on its members, granting they are responsible for the reputation of their group. Once we consider the university as our target group environment, its members would be its staff. They are responsible for doing research, and by publishing their research, they start to be noticed and receive recognition, which is naturally transmitted to their universities as well. Due to it, first, we need to rank researchers correctly so we can rank groups properly.

The research staff of a university varies, it is mostly composed of professors (and their most different levels), postdoctoral fellows, doctoral students, among others. Normally, professors are responsible for the research groups, where postdoctoral fellows and doctoral students work. At some point, the group publishes a paper and the

[1] http://acm.org/sigs
[2] https://scholar.google.com.br
[3] http://academic.microsoft.com
[4] http://aminer.org
[5] http://citeseerx.ist.psu.edu
[6] http://semanticscholar.org
[7] http://csrankings.org

responsible professor is typically involved. It allows us to use professors as the connecting factor, consequently transferring the paper reputation to the group. The indication of the author's affiliation is key to estimate reputation for groups. However, it may be uncertain when considering sub-areas due to venues and authors sub-areas.

Some esteemed venues cover multiple sub-areas and it may lead to inadequate interpretations. For instance, a sub-area such as Information Retrieval (IR) has a strong relation to Databases, meaning that Databases researchers are capable of publishing Database related papers in IR venues because some of them have a more multidisciplinary approach and can fit different papers.

One example of such venue is the ACM Conference on Information and Knowledge Management (CIKM). It is a renamed venue that works mainly in three areas: Databases, Information Retrieval, and Knowledge Management. The last two sub-areas are relatively new compared to Databases, even though they all have some sort of correlation. It becomes an issue because P-score does not differ venues based on their areas, their scores are given based on their papers. Since CIKM has an important contribution to IR and there are publications not related to IR there, we need to identify the areas of each paper and compute differently their contribution.

Another problem we face is the author's area of research itself. It is possible to infer the area of research based on one's publications, usually, they tend to publish in the same or related venues. This process requires to consult researchers history of publications, the more they publish in a certain venue, the more likely it is to assume that they do research in that area. Since this evaluation is based on the history of publications, if they decide to change their field of study, it is going to take some time to notice it using this approach.

To address the mentioned issues, we estimate the authors' relation to each sub-area and, then, use it to weight their contribution for each sub-area. To do so, we observe the portion of papers published in venues restricted to that one sub-area, which are selected according to the normalized P-score, presented in Section 3.2.

At first, we compute the normalized P-score for every venue based on a sub-area. The score represents the venue relation to that sub-area. Hence, we select all venues above a certain threshold, which are going to compose the set of venues restricted to that sub-area. As soon the set is obtained, it is possible to evaluate authors according to sub-areas.

4 EXPERIMENTAL SETUP

4.1 Academic Search Dataset

[Characterize DBLP dataset]

[Defining the CS Subareas]

4.2 Evaluation Procedure

4.3 Baselines

4.4 Subarea Authors

Table 1: The Microsoft 37 Subareas of Computer Science

CS Subareas	
Algorithm	Internet privacy
Artificial intelligence	Knowledge management
Bioinformatics	Machine learning
Cognitive science	Management science
Computational biology	Mathematical optimization
Computational science	Multimedia
Computer architecture	Natural language processing
Computer graphics	Operating system
Computer hardware	Operations research
Computer network	Parallel computing
Computer security	Pattern recognition
Computer vision	Programming language
Data mining	Real-time computing
Data science	Simulation
Database	Speech recognition
Distributed computing	Telecommunications
Embedded system	Theoretical computer science
Human-computer interaction	World Wide Web
Information retrieval	

5 EXPERIMENTAL RESULTS

5.1 Subarea Venues

IR

5.2 Subarea Groups

UMass, Delaware, Virginia

6 CONCLUSIONS

ACKNOWLEDGMENTS

This work was partially sponsored by the Brazilian National Institute of Science and Technology for the Web (MCT/CNPq 573871/2008-6) and the authors' individual grants and scholarships from CNPq and Google.

REFERENCES

[1] Fabrício Benevenuto, Alberto HF Laender, and Bruno L Alves. 2015. How Connected are the ACM SIG Communities? *SIGMOD Record* 44, 4 (2015).

[2] Sujatha Das Gollapalli, Prasenjit Mitra, and C Lee Giles. 2011. Ranking authors in digital libraries. In *Proc. of JCDL*. 251–254.

[3] Glauber D. Gonçalves, Flavio Figueiredo, Jussara M. Almeida, and Marcos André Gonçalves. 2014. Characterizing scholar popularity: A case study in the Computer Science research community. In *Proc. of JCDL*. 57–66.

[4] Apirak Hoonlor, Boleslaw K. Szymanski, and Mohammed J. Zaki. 2013. Trends in Computer Science Research. *Commun. ACM* 56, 10 (Oct. 2013), 74–83. DOI: https://doi.org/10.1145/2500892

[5] Huajing Li, Isaac Councill, Wang-Chien Lee, and C. Lee Giles. 2006. CiteSeerx: An Architecture and Web Service Design for an Academic Document Search Engine. In *Proceedings of the 15th International Conference on World Wide Web (WWW '06)*. ACM, New York, NY, USA, 883–884. DOI: https://doi.org/10.1145/1135777.1135926

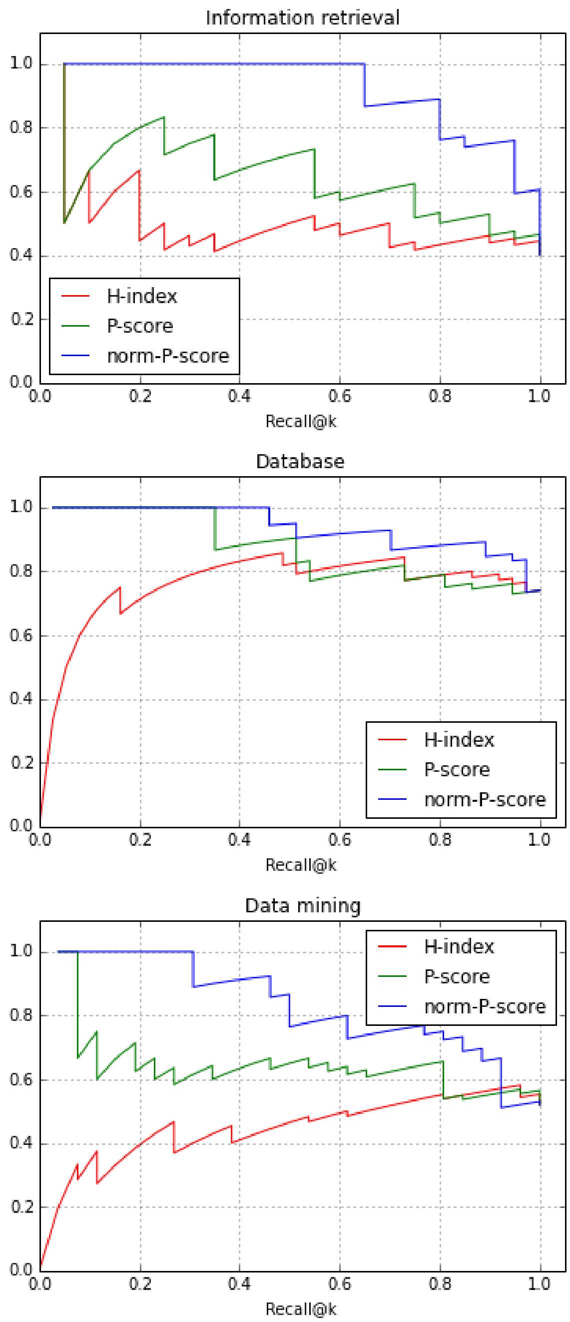

Figure 1: Precision-Recall curves of H-index, P-score and normalized P-score for the subareas of Information Retrieval, Database and Data Mining.

[6] Harlley Lima, Thiago H. P. Silva, Mirella M. Moro, Rodrygo L. T. Santos, Wagner Meira Jr., and Alberto H. F. Laender. 2013. Aggregating productivity indices for ranking researchers across multiple areas. In *Proc. of JCDL*. 97–106.

[7] Waister Silva Martins, Marcos André Gonçalves, Alberto H. F. Laender, and Gisele L. Pappa. 2009. Learning to assess the quality of scientific conferences: a case study in computer science. In *Proc. of JCDL*. 193–202.

[8] Sabir Ribas, Berthier Ribeiro-Neto, Edmundo de Souza e Silva, Alberto Hideki Ueda, and Nivio Ziviani. 2015. Using Reference Groups to Assess Academic Productivity in Computer Science. In *Proc. of WWW*. 603–608.

[9] Sabir Ribas, Berthier Ribeiro-Neto, Rodrygo LT Santos, Edmundo de Souza e Silva, Alberto Ueda, and Nivio Ziviani. 2015. Random Walks on the Reputation Graph. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. ACM, 181–190.

[10] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnet-Miner: Extraction and Mining of Academic Social Networks. In *KDD'08*. 990–998.

[11] Jacques Wainer, Michael Eckmann, Siome Goldenstein, and Anderson Rocha. 2013. How Productivity and Impact Differ Across Computer Science Subareas. *Commun. ACM* 56, 8 (Aug. 2013), 67–73. DOI: https://doi.org/10.1145/2492007.2492026