A Ranking System For Scholarly Data Analysis

Michael Shell
School of Electrical and
Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: http://www.michaelshell.org/contact.html

Homer Simpson
Twentieth Century Fox
Springfield, USA
Email: homer@thesimpsons.com

James Kirk
and Montgomery Scott
Starfleet Academy
San Francisco, California 96678–2391
Telephone: (800) 555–1212
Fax: (888) 555–1212

***Abstract*—Scholarly search engine brings great convenience in querying articles and discovering knowledge. Up to now, some systems have been developed with a tremendous amount of scholarly data and providing a range of functions to query keywords. However, these systems barely support different types ranking or various entities ranking (*e.g.,*venues, affiliation and authors) when researchers tend to get better knowledge about the keywords. In the paper, we design and develop a novel ranking system for scholarly data analysis that uses graph engine to efficiently manage the heterogeneous, evolving and dynamic natures of structured scholarly data and integrates different types ranking and various entities ranking based on SARank.**

I. Introduction

Scientific research plays a key role in promoting the development of society across the era. As a result of scientific research, scholarly articles accelerate the dissemination of scientific discoveries all around the world.

Researchers prefer to retrieve influential and recent works in massive scholarly articles that will inspire their future works. In order to do this, we need to rank articles by making full use of the involved entities (*i.e,*articles,venues and authors) and assessing the relevance of searching conditions to articles.

Generally speaking, a ranking is *a function that assigns each item a numerical score.* Scholarly articles involve with multi entities such as articles, authors, venues, affiliations and references which form a complex heterogeneous graph. Hence, scholarly articles ranking is essentially a problem of assessing the importance of nodes in a heterogeneous graph. There still remains some challenges in ranking and analysing scholarly data. Firstly, high quality heterogenous structured scholarly data is difficult to manage and operate that has exceeded one billion nodes and two billion relationships. Furthermore, academic data keeps increasing at around 5.7 million per year [1]. Secondly, even if we are only to rank one type of entities(*i.e.,* scholarly articles), the other type of entities such as venues and authors are closely involved. Moreover, the impact of different types of entities on the ranking of scholarly articles differ from each other. Finally, the importance of entities change with time in a complex manner. Recently published articles are more likely to have a increasing impacts in next few years and those published many years ago tend to have decreasing impacts since researchers potentially interested in latest discoveries.

Actually academic search engine has drawn significant attentions from both academic and industry. Some academic search engines are prevalent over the past decade *e.g.,* Google Scholar [2], Microsoft Academic [1], Semantic Scholar [3], CiteSeerX [4], Aminer [5] and AceMap [6]. Their common goal is to help researchers discover academic information and provide appropriate ways to present academic data. However, there still remains several issues in these systems. For example, when ranking papers for queries some of them focus on citation of articles [5], [6], or exploits the time-dependent information of scholarly data in the form of exponential decay(?), which fails to capture the diverse citation patterns of individual articles. Furthermore, some of them fail to support to query or rank affiliation, venue information which is also practicable to researchers.

Contributions. To this end, we build a novel ranking system for scholarly data according to SARank, to provide better rankings for researchers that realizes a range of functions including: (1) easily and efficiently search academic information by keywords; (2) rank various scholarly entities (*i.e.,* article, author, affiliation and venue) by different ranking type (*i.e.,* SARank, relevance rank, citation and year); (3) check home pages of affiliation, author, venue *etc.;* (4) find detailed information about an article.

(1) We construct a huge heterogenous academic data property graph based on graph engine that achieve great performance in querying academic data.

(2) We develop a ranking model for various entities (*i.e.,* articles, author, affiliation and venue) by evaluating the importance of the entity in academic graph. And the importance captures the temporal nature of entities.

(3) We build a prototype system for ranking and analysing scholarly data that provide better services for researchers.

Organization. The rest of our paper is organized as follows. Section 2 introduces the ranking model including author ranking, venue ranking affiliation ranking and article ranking. We give a system overview in Section 3, which a gives a brief description of our system. And the conclusion is shown in Section 4.

II. Ranking Model

ranking models in our work.

A. Author Ranking

author ranking

B. Venue Ranking

venue ranking

C. Affiliation Ranking

affiliation ranking

D. Article Ranking

article ranking

III. System Overview

system contains[7] three [1] parts.

IV. Conclusion

conclusion

References

[1] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu, and K. Wang, "An overview of microsoft academic service (mas) and applications," in *Proceedings of the 24th international conference on world wide web*. ACM, 2015, pp. 243–246.

[2] "Google scholar," https://scholar.google.com/.

[3] "Semantic scholar," https://www.semanticscholar.org/.

[4] H. Li, I. G. Councill, W.-C. Lee, and C. L. Giles, "Citeseerx: an architecture and web service design for an academic document search engine," in *Proceedings of the 15th international conference on World Wide Web*, 2006, pp. 883–884.

[5] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 990–998.

[6] Z. Tan, C. Liu, Y. Mao, Y. Guo, J. Shen, and X. Wang, "Acemap: A novel approach towards displaying relationship among academic literatures," in *WWW '16 Companion Proceedings of the 25th International Conference Companion on World Wide Web*, 2016, pp. 437–442.

[7] S. Ma, C. Gong, R. Hu, D. Luo, C. Hu, and J. Huai, "Query independent scholarly article ranking," in *ICDE 2018: 34th IEEE International Conference on Data Engineering*, 2018.