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

We first introduce Time-Weighted PageRank for evaluating the importance of entities, defined as a combination of the prestige and popularity, and then introduce entity ranking including author ranking, venue ranking, affiliation ranking and type ranking *e.g.,* relevance ranking.

A. Time-Weighted PageRank

PageRank is a typical method in scholarly articles ranking as we can easily make use of the reference between different articles. Due to the following problems, (1) xxx, (2)xxx. We introduce Time-Weighted PageRank(TWPageRank) by extending a time decay factor because of the impact of an article decay with time after peak time.

We present TWPageRank that evaluate the prestige of nodes in a directed graph, in which each node attached with time information. And we use *the impact weight* to describe the relative weight from the edge sources to targets. Formally, the impact weight on a directed edge $(u, v), i.e.$, an edge u from v, is defined as:

$$w(u, v) = \begin{cases} 1 & T_u < Peak_v \\ e^{\sigma(T_u - Peak_v)} & T_u \geq Peak_v \end{cases} \tag{1}$$

where T_u is the time of node u, $Peak_v$ is the peak time of node v after which the impact weights of edges to v decay with time, and σ is a negative number controlling the decaying speed of the impacts. By default, we use years as its time granularity in Eq. (1).

Thus, the update rule of Time-weighted PageRank is

$$PR(v) = d \sum_{(u,v) \in E} \frac{w(u, v) PR(u)}{W(u)} + \frac{1 - d}{n} \tag{2}$$

where $PR(u)$ and $PR(v)$ are the TWPageRank score of u and v. And E is a set of edges, $W(u) = \sum_v w(u, v)$ is the sum of the impact weights on all edges from u, n is the number of nodes and d is a damping parameter in (0, 1).

B. Entity Ranking and Type Ranking

Scholarly entities(*e.g.,* affiliation, venue, author and article) ranking is a problem of assessing the importance of nodes in a heterogeneous network. The importance is a combination of *prestige* and *popularity* to capture the evolving nature of entities. The prestige of scholarly entities is derived by applying TWPageRank on the citation graph G, and each type of entity is assigned the corresponding TWPageRank score as its prestige score Prs.

To learn about the popularity of different scholarly entities, we first introduce the popularity of an article. The popularity of an article v is the sum of all its citation freshness, *i.e.,* the closeness to the current year:

$$Pop(v) = \sum_{(u,v) \in E^c} e^{\sigma(T_0 - T_u)} \tag{3}$$

Here, T_0 is the current year, T_u is the largest year among all articles, σ is the negative decaying factor in Eq. (1).

Intuitively, prestige favors those with many citations soon after the publication of articles or associated articles of venues and authors, and popularity capture the temporal nature of entities.

Affiliation Ranking. We computes the importance of affiliations by their associated articles. As the importance of an affiliation evolves with time, we treat the affiliation importance in each year individually, and its importance is the sum of importance in all individual years.

We construct an affiliation graph $G^a(V^a, E^a)$ using the citation information among affiliations, in which a node represents an affiliation in a specific year and a direct edge (s, t) means that there exists articles of affiliation s citing articles of affiliation t. Thus, the impact weights are defined as sums of impact weights from affiliation s to affiliation t, *i.e.,*

$$w_a(s, t) = \sum_{u \in C(s), v \in C(t)} w(u, v) \tag{4}$$

Here, $C(s)$ and $C(t)$ are the sets of articles of affiliation s and affiliation t, and $w(u, v)$ is the impact weight of articles u and v.

The prestige of an affiliation in a specific year (Prs_a) is computed using the impact weights Eq. (4) and the update rule in Eq. (2). The popularity of an affiliation in a specific year (Pop_a) is defined as the average popularity of its articles that is computed using Eq. (3). Thus, the *affiliation importance score* (Imp_a) is defined as a combination of its prestige and popularity:

$$Imp_a(v) = Prs_a(v)^{\lambda} Pop_a(v)^{1-\lambda} \tag{5}$$

Here, $\lambda \in [0, 1]$ is the importance factor, indicates the weighting about prestige and popularity.

Venue Ranking. We computes the importance of venue using their associated articles which is similar with affiliation ranking. We treat the venue importance in each year, and construct a venue graph $G^v(V^v, E^v)$ using citation information among venues. And then we combine the prestige of a venue (Prs_v) and the popularity of venue (Pop_v) as the *venue importance score* (Imp_v).

Author Ranking. We evaluate the importance of each author, and compute the average importance of the authors of an article as its *author importance score.* However, it is obvious that the author citation graph is too large to handle. Hence, we evaluate the prestige, popularity of the author by using the average prestige, popularity of all her/his published articles, respectively. Then, the author importance score (Imp_{aut}) is the combination of its prestige and popularity, similar to affiliation ranking.

Article Ranking. If we are only to rank scholarly articles, the other type of entities such as venues and authors are closely involved. Hence, we assemble the importance of article, venue and author to produce the final scholarly articles ranking, illustrated in Fig. 1. Venue ranking and author ranking have presented in previous paragraphs. Next, introduce how to compute the importance of article using citation information.

We first construct a citation graph $G^c(V^c, E^c)$ using citation information among articles. The prestige of articles (Prs_c) is derived by using TWPageRank in citation graph G^c. The popularity of an article (Pop_c)is the sum of all its freshness which has described in Eq. (3). The importance of citation component

(Imp_c) is a combination of its prestige and popularity in the citation graph by applying Eq. (5).

Thus, the static ranking score of an article v is aggregated as follows:

$$S(v) = \alpha Imp_v(v) + \beta Imp_{aut}(v) + (1 - \alpha - \beta) Imp_c(v) \quad (6)$$

Here parameter α, β and $1 - \alpha - \beta$ regularize the contributions of the venue, author and citation information.

Relevance Ranking. We have introduced affiliation, venue, author, article ranking in the former sections. These rankings are query independent and aim to give a static ranking based on scholarly data only. However, it is vital to evaluate the similarity between the short query and sentence (*i.e.,* title of the paper) when retrieve articles by keywords.

Hence, we apply distributional semantic approach to represent words. Similarities between the term vectors indicates the corresponding semantic similarities [7]. The final ranking score of an article v that relates to semantics of article title, defined as follow:

$$F(v) = \theta S(v) + (1 - \theta) \sum_j idf(Q_j) \frac{Q_j \cdot T}{\|Q_j\| \, \|T\|} \quad (7)$$

Here, $S(v)$ is a static ranking score of article, $Q(j)$ is the jth of query keywords vector which have removed stop words, $idf(Q_j)$ is inverse document frequency measures how much information the word $Q(j)$ provides, T suggests we aggregate the title word vectors to their centroid and θ means the relevance factor.

III. System Overview

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] G. S. Corrado, J. Dean, K. Chen, and T. Mikolov, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.