Evaluating Score and Publication Similarity Functions in Digital Libraries

S. Bani-Ahmad[1], A. Cakmak[1], A. Al-Hamdani[2], and Gultekin Ozsoyoglu[1]

[1] Case Western Reserve Univ, USA
{sulieman, ali.cakmak, tekin}@case.edu
[2] Dept. of Computer Science, Sultan Qaboos University, Oman
abd@squ.edu.om

1 Introduction

Digital libraries do not assign importance/relevance scores to their publications, authors, or publication venues, even though scores are potentially useful for (a) providing comparative assessment, or "importances", of publications, authors, publication venues, (b) ranking publications returned in search outputs, and (c) using scores in locating similar publications. Using social networks and bibliometrics, one can define several score functions.

Existing publication similarity functions, used to locate similar papers to a particular paper, fall into two classes, namely, text-based similarity functions from Information Retrieval, and citation-based similarity functions based on bibliographic coupling and/or co-citation. In this study, we propose a number of publication, author, and publication venue score functions and publication similarity functions, which are then extended and evaluated in terms of accuracy, separability, and independence.

2 Experimental Setup

For each paper in *ACM SIGMOD Anthology (AnthP)*, we extracted titles, authors, publication venues, publication year info, and citations. The experimental dataset includes (a) 106 conferences, journals, and books, (b) 14,891 papers, and (c) 13,208 authors. For more details, see [1].

3 Score and Similarity Functions

Existing citation-based publication score functions are all based on the notion of prestige in social networks [2] and bibliometry [3]. As paper score functions we use (i) the well-known PageRank [4] algorithm, (ii) the *authorities* score of HITS (Hyperlink Induced Topic Search) algorithm [5], and (iii) the normalized citation count which, for paper P that receives C_P citations, is computed as the percentage of papers that receive C_P citation or less[6].

We compute author importances in four different ways. All author importance functions are computed by averaging the scores of selected papers of a given author A. For more details about different scoring functions, see [1].

E.A. Fox et al. (Eds.): ICADL 2005, LNCS 3815, pp. 483–485, 2005.

We use bibliographic coupling as a similarity indicator between papers, and propose a number of similarity measures based on (extended)bibliographic coupling similarity and considering the citations iteratively, which we refer to as *reachability analysis*. We also utilize paper scores to explore additional alternatives to compute paper similarities. Finally, we define a number of different co-citation-based similarity functions between papers.

Also, we compute similarity between two papers based on author-coupling (i) directly via the number of common authors between the two papers, and (ii) indirectly via co-authorship in other papers. For more details about different similarity functions, see [1].

4 Major Findings

Our major findings in this study are as follows:

* Among paper scoring functions, the citation-count-based scoring is the best in terms of separability. PageRank-based scoring is the best in terms of accuracy.

* Authorities scores and PageRank scores of papers are highly correlated.

* Separability of PageRank-based paper scores can be enhanced by (a) weighing citations, (b) weighing the *Future Citation Probabilities* represented by the E parameter of PageRank, (c) postprocessing PageRank raw scores by (i) nonlinear normalization, or (ii) linear normalization via a properly selected percentile score or (iii) combining PageRank-based paper scores and publication venue scores.

* Author scores based on author's top K-scored or top-K% scored papers accurately capture author scores.

* Citation-count-based publication venue scores are more accurate than author-score-averaging publication venue scores published in publication venues.

* By evaluating *multiple levels* of paper similarities based on bibliographic-coupling, co-citation and author-coupling, we observe that: (a) similarity value distribution curves are similar within the same group of similarity functions, (b) citation-based and author-coupling based similarity functions are more separable than bibliographic-coupling-based functions, (c) top-K overlapping ratio between paper similarity functions increases as we move to *higher levels* of similarity functions since more papers appear to be similar, (d) text-based similarity function show very low overlapping with citation-based and author-coupling-based functions.

References

1. Bani-Ahmad, S., Cakmak, A., Al-Hamdani, A., Ozsoyoglu, G.: Evaluating score and publication similarity functions in digital libraries. Technical report, CWRU (2005)
2. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge U. Press, Cambridge (1994)
3. Chakrabarti, S.: Mining the Web. Morgan-Kauffman, San Francisco, CA (2003)

4. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford DL Technologies Project (1998)
5. Kleinberg, J.: Authoritative sources in hyperlinked environments. In: the 9th ACM-SIAM Symposium on Discrete Mathematics. (1998)
6. Ozsoyoglu, G., Altingovde, I., Al-Hamdani, A., Ozel, S., Ulusoy, O., Ozsoyoglu, Z.: Querying web metadata: Native score management and text support in databases. ACM TODS (2005)