Award Prediction with Temporal Citation Network Analysis

Zaihan Yang Dawei Yin Brian D. Davison
Department of Computer Science and Engineering
Lehigh University
Bethlehem, PA, 18015 USA
{zay206 | day207 | davison}@cse.lehigh.edu

ABSTRACT

Each year many ACM SIG communities will recognize an outstanding researcher through an award in honor of his or her profound impact and numerous research contributions. This work is the first to investigate an automated mechanism to help in selecting future award winners. We approach the problem as a researchers' expertise ranking problem, and propose a temporal probabilistic ranking model which combines content with citation network analysis. Experimental results based on real-world citation data and historical awardees indicate that some kinds of SIG awards are well-modeled by this approach.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Performance

Keywords: citation network, link analysis, temporal correlation

1. INTRODUCTION

Each year many ACM SIG communities will recognize an outstanding researcher through an award in honor of his or her profound impact and numerous research contributions. The most recent Salton award winner (year 2009) in the SIGIR community, for example, Dr. Susan Dumais, is widely acknowledged as an IR expert due to her contributions in both theoretical development and practical implementations of Latent Semantic Indexing and question-answering. Winning such an award is thus a particularly strong indication of expertise and prestige in a given field. Even though there has been research in evaluating scientists' reputation and thus finding experts in a certain field, no work has developed an automatic and efficient mechanism in selecting future award winners. This work takes the first step into this problem.

We approach the problem as a researchers' expertise ranking problem. In one direction of the approaches in evaluating the expertise of a researcher, different information probabilistic models have been provided, including language model [1], voting model [5], and discriminative model [3], which mainly emphasize evaluating the relevance between supporting documents and thus the corresponding authors with the query. Another direction of research, which is the research focus of this poster, takes use of social network analysis [2, 8] to boost ranking performance. However, in both of these approaches, one important factor has largely been ignored by previous research: temporal information. As the awards of a SIG community are often issued annually, the authority of a researcher varies over time. In this paper, we propose a novel temporal citation network analysis model to predict SIG-award winners.

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

2. THE MODEL

Our weighted citation network can be represented as G=<A,E>, where A is a set of author nodes, and E is a set of edges. Two types of relationships (edges) between pairs of authors have been considered: coauthorship and citations. Thus, $(a_i, a_j) \in E$ if author a_i coauthored with author a_j or if at least one of the publications of author a_i cites a publication of a_j.

2.1 Temporal Factor

We introduce four temporal factors to represent an individual researcher's academic activity. 1) $CareerTime$: How long has a researcher been publishing papers? We assume that the longer the career time a researcher has, the higher authority he may have. 2) $LastRestTime$: How many years have passed since the last publication of a researcher? We assume that a long time without academic output will negatively affect a researcher's scholarly reputation. 3) $PubInterval$: How many years on average would a researcher take between every two consecutive publications? We assume that more frequent publication indicates more active academic participation. There is one other temporal factor which considers the long-lasting influence of a researcher's publication, and thus indirectly represents the influence of the researcher. We assume that if a paper continues to be cited a long time after its publication, it brings higher prestige to its author (e.g., the paper PageRank [6] is frequently and persistently cited by subsequent papers). To model this temporal factor, we first introduce a decay function to differentiate the weight between a pair of paper citations. If paper p_j published in year y_j cites another paper p_i published in year y_i $(y_j - y_i) \geq 0$, we define the *citation influence ratio* of paper p_j on p_i as: $CIR(p_{ji}) = \beta_1(1 - \beta_2^{y_j - y_i})$, where β_2 $(0 < \beta_2 < 1)$ is the decay base. We now define the *citation influence* between a pair of authors as: $CI(a_{ji}) = \sum CIR(p_{ji})$, where p_j is any paper of author a_j, p_i is any paper of a_i, and p_j cites p_i.

2.2 Temporal Authority Propagation Model

Based upon the discussion above, we define an *individual temporal importance* (ITI) to model the researcher's academic authority in terms of time. The ITI of author a_i can be expressed as: $ITI_i = CareerTime_i * (1/LastRestTime_i) * (1/PubInterval_i)$. The weight on an edge from a_i to a_j can then be defined as: $\omega(a_{ij}) = (NumCo(a_{ij}) + CI(a_{ij})) * ITI_j$, where $NumCo(a_{ij})$ is the number of times author a_i coauthored with a_j. We normalize the weights on edges over the whole network by defining the propagation probability from a_i to a_j as: $P(a_i, a_j) = \frac{\omega(a_i, a_j)}{\sum_{k:i \to k} \omega(a_i, a_k)}$. Under this definition, author a_i will propagate more authority to author a_j if they coauthored more often, if a_i has greater citation influence on a_j, or if a_j has greater individual temporal importance. Similar to the original PageRank

Table 1: Prediction performance across algorithms

	NumPub	NumCit	in-domain	in-domain	H-index	LM	CoRank	PR	TAP
			NumPub	NumCit					
NumTop10	28	30	25	31	13	2	20	30	**34**
NumTop20	38	40	38	40	18	2	28	36	**47**
MRR-All	0.1065	0.1189	0.0992	0.1183	0.0495	0.0080	0.1031	0.1053	**0.1291**

Table 2: Top20%: Individual SIG Award Prediction Results

SIGARCH	SIGSOFT	SIGPLAN	SIGKDD
0.35	0.40	0.21	0.71

SIGIR	SIGCOMM	SIGMOD	SIGCSE
0.60	0.17	0.65	0.45

[6] function, the propagation function in our model can be represented as: $PR(i) = (1-d)\sum_{j:j\rightarrow i} PR(j) * P(j,i) + d\frac{1}{N}$, where is N is the total number of author nodes.

3. EXPERIMENTAL WORK

From the ACM digital library[1], we crawled the descriptive web pages for published papers as our experimental dataset. For each publication, we extracted and recorded the information of its publishing year, authors, and citation references. We finally captured 170,897 authors and 172,890 papers. We retrieved for each year a time-based subset of all the papers and authors, which means if we aim to predict the award winners of 2009, we would first retrieve all the papers published before 2009, and their corresponding authors to build the graph.

3.1 Evaluation

In the portal website of Microsoft Academic Search[2] (a free computer science bibliography search engine), we found 23 categories covering the main disciplines of computer science research. For 6 of them, we collected the corresponding SIG awards in the ACM community. They are the awards for SIGCSE (20), SIGPLAN (19), SIGCOMM (18), SIGMOD (17), SIGARCH (17), SIGSOFT (15). We choose them because they have more examples of award winners. We furthermore collected the award winners for SIGKDD (7) and SIGIR (5) community for the sake of our interests. The number in the parenthesis indicates the number of award winners from 1990 to 2009 (our predicting period) that can be found in our dataset. As a result, we used these 8 categories as testing queries, and the 118 existing award winners as ground truth.

We further generate a profile for each author a by concatenating all of his publications in terms of title, abstract and ACM categories, and combine the citation network ranking results with the Okapi BM25 [7] ranking results as: $\lambda * rank_{BM25}(a) + (1-\lambda) * rank_{CitationNetwork}(a)$. λ is tuned between 0 and 1 to get the best outcome for each award winner. Three metrics have been used to evaluate the performance of an algorithm. 1) **NumTop10**: the total number of award winners that can be ranked within the Top 10. 2) **NumTop20**: the total number of award winners that can be ranked within the Top 20. 3) **MRR-All**: the average MRR score across all award winners.

3.2 Experimental Results

We compared our model with several existing algorithms previously used in citation network analysis work or expert-finding work. They include the ranking by 1) overall number of publications (NumPub), 2) overall number of citations (NumCit), 3) in-domain NumPub [8], 4) in-domain NumCit [8], 5) H-index [4], 6) Language-Model based approach as introduced in [1], and 7) CoRank algorithm as introduced in [9]. We also run a weighted PageRank (referred to as PR) on the network, where when compared with our Temporal Authority Propagation (referred to as TAP) model, no temporal information is considered. The weights on a edge from a_i to a_j would then be defined as: $\omega(a_{ij}) = NumCo(a_{ij}) + NumCit(a_{ij})$, where $NumCit(a_{ij})$ is the number of times author a_i cites a_j. We combined each baseline algorithm's (except 6) ranking results with the BM25 ranking results and tuned the λ to achieve the best performance for each award winner. Parameters β_1 and β_2 play important roles in our TAP model. Preliminary experiments show that the best performance of our model will be achieved when β_1 is set to 1, and β_2 is set to 0.9. As indicated in Table 1, our model can retrieve 47 award winners within Top 20, which is 39.8% of all the existing awards winners in our data set. NumCit and in-domain NumCit give the best performance in terms of NumTop20 among all non-temporal algorithms, while our algorithm improves their performance by 17.5%. We also investigated the influence of $NumCo$ and the four temporal factors and found that all were necessary to achieve the reported performance.

We are interested in finding out what fraction of all award winners in each SIG community can be ranked within Top 20. As indicated in Table 2, our model can make good predictions on several awards, such as SIGKDD, SIGMOD, and SIGIR, but comparatively worse on others, such as SIGCOMM and SIGPLAN.

Acknowledgments

This work was supported in part by a grant from the National Science Foundation under award IIS-0545875.

4. REFERENCES

[1] K. Balog, L. Azzopardi, and Rijke. Formal Models for Expert Finding in Enterprise Corpora. In *SIGIR*, pages 43–50, 2006.

[2] H. Chen, H. Shen, J. Xiong, S. Tan, and X. Cheng. Social network structure behind the mailing lists. In *TREC-14*, 2006.

[3] Y. Fang, S. Luo, and A. Mathur. Discriminative Models of Integrating Document Evidence and Document-Candidate Associations for Expert Search. In *SIGIR*, pages 683–690, 2010.

[4] J. E. Hirsch. Citation indexing: Its theory and application in science, technology, and humanities. In *Proceedings of National Academy of Sciences*. John Wiley and Sons, Inc., 2005.

[5] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *CIKM*, pages 387–396, 2006.

[6] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. In *Stanford InfoLab, Technical Report 1999-66*, 1998.

[7] S. Robertson. Overview of okapi projects. *Journal of Documentation*, 53:3–7, 1997.

[8] Z. Yang, L. Hong, and B. D. Davison. Topic-driven Multi-type Citation Network Analysis. In *RIAO*, pages 24–31, 2010.

[9] D. Zhou, S. A. Orshanskiy, H. Zha, and C. Giles. Co-ranking authors and documents in a heterogeneous network. In *ICDM*, pages 739–744, 2007.

[1] http://portal.acm.org

[2] http://academic.research.microsoft.com