2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology

Using Web-Mining for Academic Measurement and Scholar Recommendation in Expert Finding System

Chi-Jen Wu, Jen-Ming Chung, Cheng-Yu Lu*, Hahn-Ming Lee‡, and Jan-Ming Ho
Institute of Information Science, Academia Sinica, Taiwan
‡*Dep. of CSIE, National Taiwan University of Science and Technology, Taiwan*
**Corresponding Author E-mail:cylu@iis.sinica.edu.tw*

***Abstract*—Scholars usually spend great deal of time on searching and reading papers of key researchers. However, to objectively determine key researcher of a topic relies on several measurements, such as publication, citation, recent academic activities. In this paper, a prototype of scholars searching and recommendation system based on a web mining approach in expert finding system is proposed. The system gives and recommends the ranking of scholars and turns out top-k scholars. A new ranking measure is designed, namely p-index, to reveal the scholar ranking of a certain field. We use a real-world dataset to test the robustness, the experiment results show our approach outperforms other existing approaches and users are highly interested in using the system again.**

***Keywords*-Academic Measure; Web Mining; Expert Finding System;Performance Indexing;**

I. Introduction

In an expert finding system (EFS), it is required to recommend important researchers of a research topic. Generally, researchers are judged by counting his/her publications instead of considering the quality of his/her papers. However, the aspect should be switched to concern the quality of their publications. Therefore an interesting challenge arises, how to measure and recommend important/famous scholars of a research topic? In fact, constructing rankings of scholar authorities is a relatively new subfield of information retrieval research. This problem is different to the traditional expert finding problem [1] [2], in essence, the goal of EFS is to identify a list of people with relevant expertise of a topic. However, the scholar searching problem is a deeper expert finding problem, it is not only identifying right scholars who possess a required knowledge, but ranking their level of authority in the research field. Generally, how to find key researchers is more complex and difficult than finding experts; Particular, there is no standard specifying the criteria or popular qualifications necessary for particular levels of authority of scholar.

In this paper, we propose a new design of a scholar searching system prototype by using web mining approach; We first focus on the problems of scholar finding and scholar ranking. Our system assorts the ranking of scholars with relevant expertise of a research area, such as "Signal Processing", "Data Mining", and turns out top-k important scholars. For scholar finding, search engines are employed to analyze documents of a certain topic, and extract the authors from the received documents. Then we estimate the extracted author's relevance to the topic on web pages through statistical analysis. We assume that authors with a plenty of articles about a certain topic are more likely to be a candidate expert and authors with highly cited papers are indicative of the authorities. For scholar ranking, we design p-index, a ranking function for positioning scholars. The ranking criteria of scholars are based on publications, citations by computing the query results from the scientific literature digital archive, such as Google Scholar and MS Libra Academic Search. The ranking function, called p-index, is a novel measure to estimate an individual scholar's impact of a single field. The p-index is to indicate the total citation of a scholar's papers is m% percentage of total citation of whole papers in this research field.

II. Related Research

Expert finding is a task of finding right scholars of a certain topic with high relevance. Within a research community, such as computer science, there should be many possible candidates who are relevant to a given topic, the expert finding operation retrieves a list of expert candidates who are deemed the most likely scholars for this topic. Second, expert ranking [3] [4] assorts the levels of authority among the candidates, and it involves analysis of reputation, publication, citation, and activities among a list of candidate scholars. Finally, expert profiling [5] [6] to dig and extract the profile information of an individual scholar from the Web, it includes basic information, contact information, and the educational history. In this section, we describe the related work includes the above three components.

Traditional expert finding is to identify a list of people who are with appropriate skills and knowledge related a given topic [7]. Most previous approaches rely on the development of an expert database by deploying manual processes [8], or base on the text, citation or document analysis in matching user's research topic [2] [9] [10] [11].

We are aware that a few systems employed expert ranking techniques, such as Arnetminer, Libra, and CiteSeerX.

978-0-7695-4513-4/11 $26.00 © 2011 IEEE
DOI 10.1109/WI-IAT.2011.137

The main idea of Arnetminer is similar to Referred Web, and other systems are based on the Information Retrieval schemes. Because there is no standard specifying the criteria for particular levels of authority of scholar, to rank scholar become difficult and it is hard to result in a unanimous solution. The well-known ranking index is Impact Factor that is defined as the average number of citations per journal over a two years period. Since 2005, the *h-index* [4] has been proposed to measure an individual scholar's impact. The *h-index* indicates that a scholar has published *h* papers and these papers has been cited more than *h* times. In the state of the art research, Ren and Tayloar [3] provided an automatic publication-ranking based framework to support such ranking for scholars and research institutions. They discussed the most important ranking policy and indicated several limitations for publication-ranking.

Another important challenge of scholar searching system is expert profiling task. Specifically, it focuses on studying how to extract the profile of an individual researcher from the Web automatically [12]. Recently, Tang et al. [5] [6] present a unified approach to extract scholar profiles on an academic social network. This system also addresses the name disambiguation problem [13]. Actually, many profile extraction methods have been proposed, an overview can be found in [14].

III. Approach

In this section, we describe our scholar searching system and its components, and demonstrate the system by several experiments to recommend important researchers. First, we give an overview of the system's main concepts, the corresponding task components, and their interplay. Then we construct the system prototype based on these concepts and started experimenting, we demonstrated a number of search experiments.

A. System Overview

Our scholar searching system consists of three main components. We firstly proposed a customized crawler to collect papers from digital archiver as the candidates set, called *c* set. The crawler collects scientific literatures of a given the query topic *q* and extracts the author name information from these articles. In addition, the crawler also analyzes the citations of each literature and candidate. After obtaining *c* set, our system is to estimate the associations between a topic and candidates. For estimating the relevance to the given topic, we have the following claim.

Claim 1: Authors with a plenty of articles about a certain topic are more likely to be an expert on the topic q.

This claim should be reasonable because an important scholar, his/her name should be popular on the Web pages [11]. Because our idea is based on this claim, we estimate the extracted author's relevance to a given topic on web pages using statistical analysis. A number of statistical analysis methods are proposed for estimating term association based on co-occurrence measures [15]. In our study, Chi-square test is adopted because the required parameters for it could be easily gathered by using search engine. Then, we rank candidates according to the results of Chi-square test, and determine top-*k* candidates in the *c* set. Finally, a ranking measure, called p-index, which is a novel method to estimate an individual scholar's impact of a research field.

B. Relevance Estimation

Because of the effect and efficiency of implementation, Chi-square test is applied to estimate the strength of relation between extracted scholar and given research topic by co-occurrence of from web pages. By giving a query topic q and a candidate's name c, and we assume q and c are independent; the details of our implementation of Chi-square are referred from [16]. This Chi-square test is important as a co-occurrence index in our system. In practice, we use Google as a search engine, but other search engines are also applicable (i.e., Yahoo, Bing). The Chi-square test method provides an efficient way to estimate the relevance between candidate and research topic, and it is easy to implement in our EFS [17].

$$
\begin{aligned}
E(q,c) &= ((a+c)(a+b))/n,\\
E(q,^{\Lambda}c) &= ((b+d)(a+b))/n,\\
E(^{\Lambda}q,c) &= ((a+c)(c+d))/n,\\
E(^{\Lambda}q,^{\Lambda}c) &= ((b+d)(c+d))/n,
\end{aligned}
$$

Then, we have a conventional Chi-square test as follow:

$$
\begin{aligned}
\chi^2(q,c) &= \sum_{\forall X\in\{q,\neg q\},\forall Y\in\{c,\neg c\}} \frac{[n(X,Y)-E(X,Y)]^2}{E(X,Y)}\\
&= \frac{n\times(a\times d-b\times c)^2}{(a+b)\times(a+c)\times(b+d)\times(c+d)}
\end{aligned}
$$

C. Scholar Ranking

Existing ranking indexes have several limitations. One major limitation is that they are used to rank the whole research field, such as all of computer science. It is hard to infer the contributions of a scholar in a sub research field. Hence we design a novel ranking measure, called p-index, to estimate an individual scholar's impact. The p-index indicates the total citation of a scholar's papers is m% percentage of total citation of whole papers in this research field. We define p-index as follows:

$$
C_i^{p-index} = \frac{\sum citations \in C_i}{\sum citations \in \theta} \times 100,
$$

where C_i dedicates a scholar in c set, and θ indicates a set of collected scientific literatures. p-index could be used alone, but it should probably serve as one quantitative

Table I
THE REQUIRED PARAMETERS FOR CHI-SQUARE TEST (WE SET n=8 BILLION IN OUR EXPERIMENTS)

The required parameter	Notation
The total number of Web pages	n
The number of Web pages containing both candidate's name and topic	a
The number of Web pages containing topic but without candidate's name	b
The number of Web pages containing candidate's name but without topic	c
The number of Web pages without both candidate's name and topic	$d(d = n - a - b - c)$

Table II
THE RANKING RESULT OF DATA MINING RESEARCH AREA

Ranking	Candidate	p-index	χ^2test
1	J Han	9.6475	2652383
2	M Kamber	4.3345	3927792
3	E Frank	3.8335	2044977
4	IH Witten	3.8335	1977850
5	G Piatetsky-Shapiro	3.7289	28694484
6	P Smyth	3.5951	4588678
7	T Hastie	3.4983	2951988
8	R Tibshirani	3.4983	2219211
9	J Friedman	3.4983	186502
10	JC Bezdek	3.2806	424565

indicator in a more comprehensive methodology. In addition to publications, many important factors, such as research impact, funding, students, can reflect the importance of a scholar can be considered in future works.

Figure 2 and Table II shows the output of querying "Data Mining" and retrieving top-*k* scholars in the given topic (here we set k =10)

IV. EXPERIMENT RESULTS

To validate our system, we use it to perform two rankings. The first ranking assessed the scholars in Data Mining area. First we give the perspective statistics of these two fields. Due to the limitations, our crawler only retrieves first 1,000 papers from Google Scholar.

Figure 1 depicts the citation distribution of collected papers. In this figure, we can know that top 100 papers dominate the citation impact. And there is a very high citation paper in Data mining field. In fact, it is a book, title "Neural networks: a comprehensive foundation" by Simon Haykin, has been cited 22,363 times. Although this author has a very high citation paper, his importance may not be more than top-*k* scholar in data mining field, his chi-square test is 107836 $\ll$ 180509 (JA Hartigan's chi-square test). Figure 1 also shows the first 0.1% papers dominated 95% citations in both fields. Figure 2 depicts the citation distribution of scholars. It also shows a fewer people received a great citations, that is similar to result of Figure 1.

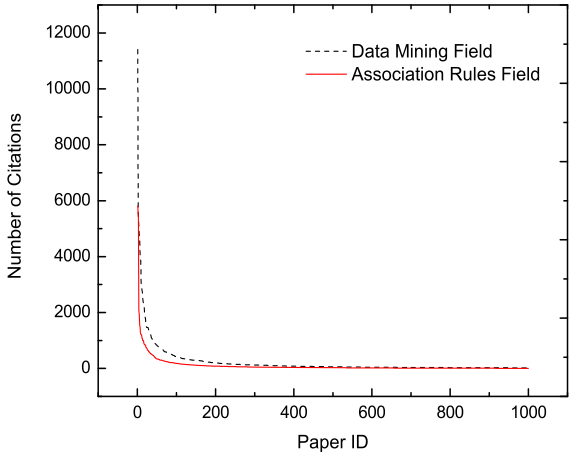

Figure 1. Citation distribution (paper).

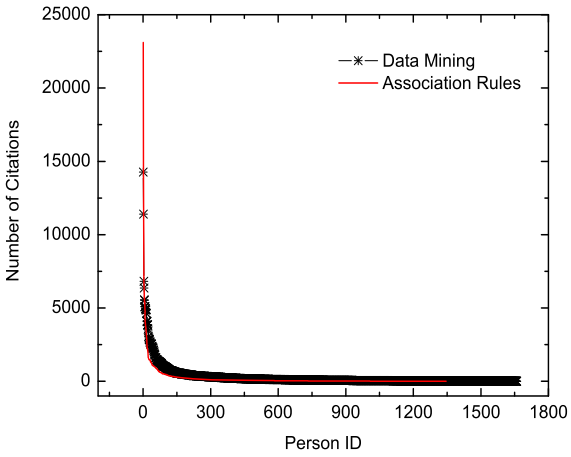

Figure 2. Citation distribution (scholar).

V. Conclusion and Future Work

In this paper, we proposed and implemented a scholar searching system prototype based on a web mining approach in our EFS [17] [18]. The system computes the ranking of scholars with relevant expertise of a topic, e.g., Data mining , and turns out top-k scholars. A new ranking function, p-index, is designed as a measure to recommend scholar in a specific research field. Our contributions include: 1) proposal of a web mining approach to famous and important scholar searching, 2) we have developed a flexible ranking function, p-index, for scholar ranking in a research field, and 3) we have constructed and demonstrated our scholar searching mechanism in our EFS. A main advantage of our approach is that users can query any research topic and find a list of authoritative and important scholars without dedicated databases for the demand.

In the future, we will firstly focus on Name Ambiguity issues, we have published some research results to improve the robustness [19]; Secondly, the proposed academic measure, p-index, will be incorporated in our *Digital Library Connector* (DLC) which provides scholars facilities to manage publication list, to subscribe important researchers' academic activities, and several recommendation services. (DLC is accessible at http://dlc.iis.sinica.edu.tw)

Acknowledgments

This work is partly supported by the National Science Council of Taiwan, under grant NSC 98-2221-E-001-010-MY3 and NSC99-2221-E-011-075-MY3.

References

[1] H. Kautz, B. Selman, and M. Shah, "Referral web: combining social networks and collaborative filtering," *Communications of the ACM*, vol. 40, no. 3, pp. 63–65, 1997.

[2] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, and A. van den Bosch, "Broad expertise retrieval in sparse data environments," *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 551–558, 2007.

[3] J. Ren and R. N. Taylor, "Automatic and versatile publications ranking for research institutions and scholars," *Communications of the ACM*, vol. 50, no. 6, pp. 81–85, 2007.

[4] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 46, p. 16569, 2005.

[5] J. Tang, D. Zhang, and L. Yao, "Social network extraction of academic researchers," *icdm*, pp. 292–301, 2007.

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

[7] N. Craswell, A. de Vries, and I. Soboroff, "Overview of the trec-2005 enterprise track," *TREC 2005 Conference Notebook*, pp. 199–205, 2005.

[8] T. Hofmann, "Probabilistic latent semantic analysis," *Proc. of Uncertainty in Artificial Intelligence*, pp. 289–296, 1999.

[9] T. Bogers, K. Kox, and A. van den Bosch, "Using citation analysis for finding experts in workgroups," *Proc. DIR*, 2008.

[10] T. Reichling, M. Veith, and V. Wulf, "Expert recommender: Designing for a network organization," *Computer Supported Cooperative Work (CSCW)*, vol. 16, no. 4, pp. 431–465, 2007.

[11] M. Harada, S. Sato, and K. Kazama, "Finding authoritative people from the web," *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pp. 306–313, 2004.

[12] C.-Y. Lu, S.-H. Lin, J.-C. Liu, S. Cruz-Lara, and J.-S. Hong, "Automatic event-level textual emotion sensing using mutual action histogram between entities," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1643 – 1653, 2010.

[13] R. Bekkerman and A. McCallum, "Disambiguating web appearances of people in a social network," *Proceedings of the 14th international conference on World Wide Web*, pp. 463–470, 2005.

[14] J. Tang, M. Hong, D. Zhang, B. Liang, J. Li *et al.*, "Information extraction: Methodologies and applications," *Emerging Technologies of Text Mining: Techniques and Applications*, 2007.

[15] R. Rapp, "Automatic identification of word translations from unrelated english and german corpora," *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 519–526, 1999.

[16] P. J. Cheng, J. W. Teng, R. C. Chen, J. H. Wang, W. H. Lu, and L. F. Chien, "Translating unknown queries with web corpora for cross-language information retrieval," *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 146–153, 2004.

[17] K. H. Yang, C. Y. Chen, H. M. Lee, and J. M. Ho, "EFS: expert finding system based on wikipedia link pattern analysis," *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*, pp. 631–635, 2008.

[18] K. H. Yang, T. L. Kuo, H. M. Lee, and J. M. Ho, "A reviewer recommendation system based on collaborative intelligence," *2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 564–567, 2009.

[19] K. Yang, H. Peng, J. Jiang, H. Lee, and J. Ho, "Author name disambiguation for citations using topic and web correlation," *Research and Advanced Technology for Digital Libraries*, pp. 185–196, 2008.