Citation Analysis for Academic Expert Finding

ABSTRACT

Bla bla bla

Keywords

citation analysis, expert finding

1. INTRODUCTION

There is an increasing belief that enterprise search is a vital tool for meeting the demands of the global marketplace. *Expert search* is considered a crucial component of an effective enterprise search system. A successful expert search system helps an organization address two important tasks, as signaled by Maybury [21]: *expert finding* and *expert profiling*. Expert finding is the task of locating individuals or communities knowledgeable about a specific topic. A complete and up-to-date overview of the experts related to a topic, task, or assignment can for instance aid an organization in rapidly recruiting an operational team to respond to a new market opportunity or threat. Expert finding involves analyzing communications, publications, and activities. It should also include the ability to rank them on multiple dimensions such as qualifications, availability, experience, and reputation.

The term expert profiling, first coined by Balog et al. [2], encompasses all activities related to assessing expertise, such as classifying and quantifying individual expertise and the expertise of entire organizational units, and validating the breadth and depth of that expertise. A successful expert profiling system would also allow organizations to identify changes in expert profiles of individuals and organizational units [21].

In general, three different sources of information for expertise attribution can be identified within organizations [21]:

- Content-based evidence is one of the most prevalently used sources of expertise in expert finding research, typically including documents and e-mails authored by employees. Homepages, resumes, and shared folders in a file system can also be used as content-based evidence of expertise.
- Organizations are made up of a variety of social networks. We assume that people who interact are likely to share expertise. Evidence of these interactions can be found in the organization structure, but also in e-mail flow, usage of software libraries, and bibliographic information. Records of information exchange in these networks provide evidence of expertise.
- A third type of evidence for expertise is activity-based: how much time did an employee spend on a project, and what are the search and publication histories of employees.

In this paper we focus on the problem of expert finding. In particular, we investigate the impact of combining two different sources of expertise—content-based and social networks—on expert finding in an average-sized academic workgroup. The research output of such a workgroup provides a stream of content-based evidence in the form of papers and technical reports. In addition, we can also benefit from the rigorous citation culture in academia. The network of citations between papers and authors is representative of the underlying social-academic network between researchers. We assume that highly cited papers are indicative of the expertise of its authors on the topics covered by those papers. We investigate the combination of this evidence with content-based methods. To our knowledge, citation analysis and content-based expert finding techniques have not yet been compared and combined; this is the contribution of the current paper.

The remainder of this paper is organized as follows. ...

2. RELATED WORK

2.1 Expert finding

Early large-scale approaches to expert finding came in the form of constructing and querying databases containing representations of the knowledge and skills of an organization's workforce. These systems tended to delegate the responsibility and workload to the employees, giving them the task to create and maintain adequate descriptions of their own continuously changing skills [21].

This disadvantage prompted a shift to expert finding techniques more supportive of the natural expertise location process [22], and more automatic approaches to expert finding such as the one by Campbell et al. (2003). They performed expert finding on e-mail collections of two different organizations, comparing a content-based approach with a graph-based augmented approach and reporting that the latter outperformed the purely content-based approach [7].

Arguably, the key development boost for the field of expert finding and expert profiling has been the introduction of the Enterprise track in TREC 2005. From its inception the track included an Expert Finding task, that triggered rapid advances in the field of expertise retrieval, in terms of modeling, algorithms, and evaluation methods.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$10.00.

Participants in the 2005 and 2006 TREC Enterprise tracks validated their work using the W3C test collection, a 2004 crawl of the World Wide Web Consortium website [9]. This collection—330,037 documents, adding up to 5.7GB, with a list of 1,092 candidate experts—contains not only web pages but also numerous mailing lists, technical documents, and other kinds of data that represent the day-to-day operation of the W3C. For the Enterprise track of TREC 2007, a new test collection was used: the CSIRO collection with 370,715 documents, totaling 4.2 GB, with a list of 3,678 candidate experts [1]. Other collections representing different types of organizations have been created, such as the UvT Expert Collection [3]. The work described in this paper is performed on a small subset of this collection (see Section **??** for more details).

Expert finding—identifying a list of people who are knowledgeable about a given topic—is usually approached by uncovering salient associations between people and topics [9]. The co-occurrence of a person with topics in the same context is commonly assumed to be evidence of expertise of that person on those topics. The majority of expert finding approaches can be divided into either *document-centric* or *candidate-centric* approaches. In the candidate-centric approach to expert finding, each expert is represented by a profile that is constructed from the expertise evidence associated with that expert. A simple way of doing this would be concatenating all documents associated with an expert into a single profile document for that expert. In a document-centric approach, the first step is retrieving documents or other forms of expertise evidence relevant for the query and then associating those retrieved documents with the different experts. Many different retrieval models have been used for both expert finding methods, as well as many extensions to existing retrieval models originally developed for common document retrieval, such as (pseudo-)relevance feedback, query expansion, using passage-level evidence, and re-ranking using static rankings [9, 27].

Approaches that combine different forms of evidence—such as using static rankings for re-ranking purposes—are especially interesting with regard to the topic of this paper. To our knowledge, the first to do so were the aforementioned Campbell et al. when they found that a graph-based approach performed better than a pure content-based approach [7]. Chen et al. (2006) took a similar approach while investigating social networks found in the mailing lists in the W3C corpus [8]. They used PageRank [23] to rank experts on centrality, and a revised version of the HITS algorithm [18] for submitting their runs. They compared this with a two-stage model that combined relevance with co-occurrence, and found that HITS performed significantly worse. They explain that the root cause for the lack of success is the specific nature of mailing list networks, which allow for reciprocal links to be added to the network much easier that the typical web link network, or citation network. Kolla et al. (2006) used a similar HITS-based re-ranking approach and reported marginal but insignificant improvements [19]. Bao et al. (2006) achieved similar results by using PageRank [4]. In contrast, the approach taken by Zhu et al. (2006) to use Google rankings turned out to be an ineffective way of improving performance [31]. Serdyukov et al. (2007) modeled the search for experts as a multi-step propagation of relevance through a hyperlinked network of relevant documents and found improvements over a one-step model [26]. Outside of TREC, another effort to use network analysis for expert location was made by Zhang et al. (2007), who used a set of network-based ranking algorithms, including PageRank and HITS, to identify expert users of a Web-based programming community [30]. They found these algorithms did not outperform simpler algorithms for expert finding.

In sum, earlier work on static rankings does not seem to yield a satisfactory answer to the question whether using static ranking techniques such as HITS and PageRank helps or hurts expert finding performance. A possible reason might be that the networks that were analyzed—mailing lists and intranet pages—are not related (enough) to expertise; they may lack uniform and overt signs of the expertise of the individuals posting emails or adding web pages.

A related research topic that shares many similarities with expert finding is automatically routing submitted papers to reviewers in conferences [5, 10, 11, 29]. All of these approaches use the sets of papers written by the individual reviewers as content-based expertise evidence for those reviewers to match them to submitted papers. The most extensive work was done Yarowksy et al., who performed their experiments on the papers submitted to the ACL'99 conference [29]. They compared both content-based and citation-based evidence for allocating reviewers and found that combining both types resulted in the best performance.

2.2 Citation analysis

Citation analysis involves assessing the research performance of individual scholars, scholarly journals, and research groups, departments, and institutions. Analyzing bibliographic networks has a rich history: the first citation indexes were developed by Eugene Garfield in the 1950s. Garfield (1979) also pioneered the use of these indexes in assessing the popularity and impact of specific articles, authors, and publications [13].

As mentioned in the previous section, we assume the degree to which a paper (or a set of papers about a topic) is cited, to be a good indicator of expertise. We are therefore interested in bibliometric indicators that help to identify the important elements in a citation network, more specifically, well-cited papers and authors. The classic example of such a bibliometric indicator is the so-called *impact factor*. Pioneered by Garfield's Institute for Scientific Information in the 1960s, the impact factor was meant to be an objective measure of the reputability of a journal [13].It is defined as the average number of citations—or average *indegree*—per article a journal receives over a two-year period.

The original impact factor formulation does not distinguish between citations: citations from journals with a high impact have the same weight as citations from low impact journals. Pinski et al. (1976) were the first to suggest a recursive impact factor to remedy this [24], with several others proposing related approaches, such as Bollen et al. (2006) who proposed using the PageRank algorithm [6, 23]. Examples of journal rankings using the PageRank algorithm can be found, for instance, on the Eigenfactor.org website[1].

In our expert finding situation we focus on a single workgroup. We therefore only cover a subset of the citation network of the workgroup's research field. We do not have the impact factors for every journal and conference proceedings. Lacking these for now, we use the indegree count for each document and author to calculate the importance of authors and documents in the network. In addition, we wish to use PageRank as a way of calculating a recursive impact factor.

Garfield, among others, has warned against using impact factors to measure the productivity of individual scientists, arguing that different scholarly disciplines can have very different publication and citation practices and that there is "wide variation from article to article within a single journal" [14]. However, we believe that the homogeneous research focus of our evaluated workgroup alleviates this problem to some extent. Furthermore, using a recursive algorithm for calculating the impact factor—such as PageRank—can also help alleviate this. We therefore decided to use these two bibliometric indicators to determine the importance of documents

[1] http://www.eigenfactor.org/

and authors: standard citation indegree, and PageRank scores.

Another measure that has been proposed as a way of estimating an individual researcher's impact is the so-called *Hirsch number* (or *h*-index) [16]. A scholar has a Hirsch number of *h* if he has published *h* papers that have been cited *h* times or more. We could not test this measure as an expertise estimator because it is better at distinguishing between scientists within an entire field than within a workgroup; we do not have access to the full network of the research fields.

There is some related work at the intersection of information retrieval and citation analysis. One of the first investigations into the usefulness of citations for document retrieval was performed by Salton (1963), who found a significant correlation between text-based document similarity and citation overlap similarity between documents [25]. Another obvious related example is the PageRank algorithm for ranking web pages, developed by Page et al. (1998), inspired by ideas from citation analysis. It has been successfully used to improve Web retrieval performance, for instance, by producing document priors or re-ranking retrieval results [23]. Some specific search engines for scholarly literature have been developed, most notably Google Scholar [17] and CiteSeer [15]. In general, such specific search engines perform 'normal' document retrieval, re-ranking the results by indegree (citation) count.

Drawing from the principle of polyrepresentation, Larsen combined text-based retrieval techniques with citation analysis, but found no significant improvements over a bag-of-words baseline [20]. More recently, Strohman et al. (2007) tested many seemingly useful measures descriptive of the citation network, but found that only combining text-based retrieval with the graph-based Katz measure significantly improved performance [28]. Finally, Fujii (2007) also combined text-based patent retrieval with the PageRank probabilities of citations between patents and found small but significant improvements in recall [12]. Overall, there seems to be a tendency for citation analysis to yield small improvements over normal text-based retrieval approaches.

3. METHODOLOGY

Bla bla bla

4. DISCUSSION & CONCLUSIONS

Bla bla bla

5. REFERENCES

[1] P. Bailey, N. Craswell, I. Soboroff, and A. P. de Vries. The CSIRO Enterprise Search Test Collection. *ACM SIGIR Forum*, 41(2):42–45, December 2007.

[2] K. Balog and M. de Rijke. Determining Expert Profiles (With an Application to Expert Finding). In *IJCAI '07: Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, pages 2657–2662, 2007.

[3] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, and A. van den Bosch. Broad Expertise Retrieval in Sparse Data Environments. In C. L. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. P. de Vries, editors, *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 551–558, New York, NY, July 2007. ACM.

[4] S. Bao, H. Duan, Q. Zhou, M. Xiong, Y. Cao, and Y. Yu. Research on Expert Search at Enterprise Track of TREC 2006. In *TREC 2006 Working Notes*, November 2006.

[5] H. Biswas and M. Hasan. Using Publications and Domain Knowledge to Build Research Profiles: An Application in Automatic Reviewer Assignment. In *Proceedings of the 2007 International Conference on Information and Communication Technology (ICICT'07)*, pages 82–86, 2007.

[6] J. Bollen, M. A. Rodriguez, and H. Van de Sompel. Journal Status. *Scientometrics*, 69(3):669–687, 2006.

[7] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise Identification using Email Communications. In *CIKM '03: Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pages 528–531, New Orleans, LA, 2003.

[8] H. Chen, H. Shen, J. Xiong, S. Tan, and X. Cheng. Social Network Structure behind the Mailing Lists: ICT-IIIS at TREC 2006 Expert Finding Track. In *TREC 2006 Working Notes*, November 2006.

[9] N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC 2005 Enterprise Track. In *TREC 2005 Working Notes*, November 2005.

[10] S. T. Dumais and J. Nielsen. Automating the Assignment of Submitted Manuscripts to Reviewers. In *SIGIR '92: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 233–244, New York, NY, USA, 1992. ACM.

[11] S. Ferilli, N. Di Mauro, T. Basile, F. Esposito, and M. Biba. Automatic Topics Identification for Reviewer Assignment. *Advances in Applied Artificial Intelligence*, pages 721–730, 2006.

[12] A. Fujii. Enhancing Patent Retrieval by Citation Analysis. In *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 793–794, New York, NY, USA, 2007. ACM.

[13] E. Garfield. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. John Wiley & Sons, Inc., New York, NY, USA, 1979.

[14] E. Garfield. Der Impact Faktor und seine richtige Anwendung. *Der Unfallchirurg*, 101(6):413–414, June 1998.

[15] C. L. Giles, K. Bollacker, and S. Lawrence. CiteSeer: An Automatic Citation Indexing System. In I. Witten, R. Akscyn, and F. M. Shipman III, editors, *DL '98: Proceedings of the Third ACM Conference on Digital Libraries*, pages 89–98, New York, NY, June 1998. ACM.

[16] J. E. Hirsch. An Index to Quantify an Individual's Scientific Research Output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, 2005.

[17] P. Jacsó. Google Scholar: the Pros and the Cons. *Online Information Review*, 29(2):208–214, 2005.

[18] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, September 1999.

[19] M. Kolla and O. Vechtomova. In Enterprise Search: Methods to Identify Argumentative Discussions and to find Topical Experts. In *TREC 2006 Working Notes*, November 2006.

[20] B. Larsen. *References and Citations in Automatic Indexing and Retrieval Systems - Experiments with the Boomerang Effect*. PhD thesis, Royal School of Library and Information Science, Denmark, 2004.

[21] M. Maybury. Expert Finding Systems. Technical Report MTR 06B000040, MITRE Corporation, 2006.

[22] D. W. McDonald. *Supporting Nuance in Groupware Design: Moving from Naturalistic Expertise Location to Expertise Recommendation*. PhD thesis, University of California, Irvine, 2000.

[23] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical

report, Stanford Digital Library Technologies Project, 1998.

[24] G. Pinski and F. Narin. Citation Influence for Journal Aggregates of Scientific Publications: Theory with Application to Literature of Physics. *Information Processing & Management*, 12(5):297–312, 1976.

[25] G. Salton. Associative Document Retrieval Techniques using Bibliographic Information. *Journal of the ACM*, 10(4):440–457, 1963.

[26] P. Serdyukov, H. Rode, and D. Hiemstra. University of Twente at the TREC 2007 Enterprise Track: Modeling Relevance Propagation for the Expert Search Task. In *TREC 2007 Working Notes*, November 2007.

[27] I. Soboroff, A. P. de Vries, and N. Craswell. Overview of the TREC 2006 Enterprise Track. In *TREC 2006 Working Notes*, November 2006.

[28] T. Strohman, W. B. Croft, and D. Jensen. Recommending Citations for Academic Papers. In *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 705–706, New York, NY, 2007. ACM.

[29] D. Yarowsky and R. Florian. Taking the Load off the Conference Chairs: Towards a Digital Paper-Routing Assistant. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, pages 220–230, 1999.

[30] J. Zhang, M. S. Ackerman, and L. A. Adamic. Expertise Networks in Online Communities: Structure and Algorithms. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, pages 221–230, 2007.

[31] J. Zhu, D. Song, S. Rüger, M. Eisenstadt, and E. Motta. The Open University at TREC 2006 Enterprise Track Expert Search Task. In *TREC 2006 Working Notes*, November 2006.