Topic-driven Multi-type Citation Network Analysis

Zaihan Yang   Liangjie Hong   Brian D. Davison
Dept. of Computer Science and Engineering, Lehigh University
Bethlehem, PA, 18015 USA
{zay206, lih307, davison}@cse.lehigh.edu

ABSTRACT

In every scientific field, automated citation analysis enables the estimation of importance or reputation of publications and authors. In this paper, we focus on the task of ranking authors. Although previous work has used content-based approaches or citation network link analyses, the combination of the two with topical link analyses is unexplored. Moreover, previous citation analysis applications are typically limited to a graph based on author citations, or a bipartite graph based on author and paper citations. We present in this paper a novel integrated probabilistic model which combines a content-based approach with a multi-type citation network which integrates citations among papers, authors, affiliations and publishing venues in a single model. We further introduce the application of Topical PageRank into citation network link analysis due to the fact that researchers may be experts in different scientific domains. Finally, we describe a heterogenous link analysis of the citation network, exploring the impact of weighting various factors. Comparative experimental results based on data extracted from the ACM digital library show that 1) the multi-type citation graph works better than citation graphs integrating fewer types of entities, 2) the use of Topical PageRank can further improve performance, and 3) Heterogenous PageRank with parameter tuning can work even better than Topical PageRank.

Categories and Subject Descriptors

H.4 [**Information Retrieval**]: Social Network; D.2.8 [**Citation Analysis**]: Graph MiningLink Analysis

General Terms

Information Retrieval, Social Network

Keywords

Citation Network, Graph Mining, Link Analysis

1. INTRODUCTION

Estimating researchers' contribution or reputation is of great importance since it can offer support when making decisions about researchers' job promotion, project funding approval, and scientific award assignments. With the rapid development of academic digital libraries, the increasing volume of online scientific literature provides abundant sources of reputation evidence in terms of researchers' (authors') publications, as well as the citation relationships among these publications, both of which can be taken advantage of in evaluating researchers' reputations.

In order to evaluate the reputation of a researcher, especially within one scientific domain, there are typically two basic approaches. One is called the content-based approach, in which relevant documents representing expertise of a researcher can be considered, and information retrieval models can be applied to evaluate the relevance between these documents and thus authors with the query topic [9, 10, 11]. Researchers' publications in the academic digital library provides such good expertise resources.

Another important approach, which is also our main focus in this paper, is via the social network analysis approach [21]. The citation network is one form of social network in which scientific factors, like authors and papers, can be represented as nodes, and their mutual interactions, like co-authorship and citation, can be modeled as edges.

Citation network analysis has long been a popular mechanism to evaluate the importance of publications and authors. Initially, citation analysis mainly focused on counting the number of citations [2, 3]. Under this scheme, an author will have higher reputation if he can be cited by many other authors.

With the recent success of graph-theoretic approaches in ranking network entities, researchers have begun to introduce link analysis approaches like PageRank [20] and HITS [7] into citation network analysis. Further attention has also been paid to integrate different kinds of citation network, including coauthor network for authors and citation reference network for papers and take advantage of their mutual reinformance to improve reputation ranking performance. The assumption in this group of approaches is that more influential authors are more likely to produce high quality and thus highly cited papers, and well-cited papers can bring greater acknowledgments to their authors.

In spite of the constant improvement in citation network analysis, including combination with content-based approach, integration of different kind of citation works, there still remain some limitations. For example, the current citation network analysis seldom goes beyond that of the citation relationship among authors or papers. PopRank [12] integrates conferences and journals, yet there are still some other useful and easily available information in the scientific literature, such as authors' affiliations. In this paper, we propose a novel probabilistic model which can integrate the citation between authors, papers, affiliations and publishing venues in a single model. To our belief, affiliation offers a good indication of authors' expertise, since high quality organizations tend to hire researchers (authors) with higher reputation.

In order to explore on the different impact among factors, we propose a heterogeneous PageRank, permitting us to consider different propagation rates among factors. Furthermore, one distinguished contribution of our work is that we introduce the topical link analysis, which has shown success in web page authority evaluation, into citation network analysis. In summary, our main contributions include:

1. Proposing a novel probabilistic model which combines content-based analysis with a multi-type citation network, integrating relationships of authors, papers, affiliations and publishing venues in one model. This model can be extended to include more types of social factors.

2. Proposing a heterogeneous PageRank random surfer model compared with the original uniform PageRank model, to reflect the impact among different factors.

3. Introducing topical link analysis into citation network analysis. In particular, Topical PageRank [18] is adopted for citation link analysis.

4. A comparative study using ACM digital library data on various PageRank extensions as well as different complexity of citation networks.

We review related work in Section 2 on the content-based and citation network analysis based approach in evaluating authors' reputations. In Section 3, we will introduce our multi-type citation network framework. Section 4 introduces the modified heterogeneous PageRank random surfer model. Section 5 introduces topical link analysis model. Experiments and results analysis will be described in Section 5. We conclude the paper in section 6.

2. RELATED WORK

Citation analysis has a long history in assessing the research performance of individual scholar, publishing journals or papers, as well as research groups. Originally, citation analysis focused on counting the number of citations. Journal impact factor [2, 3], the most classical citation indicator, is defined as the average number of citations per article a journal receives over a two-year period. Hirsch number (h-index) [6], another famous citation indicator, is also defined in terms of citation counts.

Inspired by the success of graph-theoretic approaches in ranking network entities, scientists gradually realized that simply counting the number of citation cannot represent well the true prestige. Without distinguishing between citations, the citation from a good paper with high impact will have the same weight as citations with lower impact. Pinski [5] was the first person who realized this problem and proposed an improved recursive approach. With the great success of link analysis approach, like PageRank and HITS in ranking web pages' authorities, much recent research work, such as that by Chen et al. [19], has introduced the PageRank algorithm into citation network analysis replacing hyperlinks with citation references.

Further research work has been carried out in combining the content-based approach with citation network for reputation evaluation. P. Glenisson et al. [4] did research in combining full text and bibliometric information in mapping scientific disciplines, and Bogers et al. [22] made the first investigation into combining and comparing the citation analysis with content-based approach for finding academic experts.

Research work has been carried beyond the citation network analysis domain in integrating different types of entities. Davison [1] proposed a unified approach to analyze multiple term and document relationships. With similar idea, a so-called link-fusion [13] unified link analysis framework has been proposed which considered both the intra- and inter- type link structure among multiple-type inter-related data objects. Most recently, Guan [14] proposed a multi-type framework integrating users, documents and tags for tag recommendation. In [24], Wang et al. proposed a more general and fundamental method for analyzing semantic relation among any multiple type of data. Compared to these work mentioned above,our emphasis in this paper is using multi-type factors integration for citation network analysis.

Similar to those work in web or data management research domain, researchers have already started to pay attention to the integration of different kinds of citation networks. The assumption is that different citation relationship can mutually reinforce each other, and thus can improve ranking performance. Zhou et al. [25] is a representative work in this direction in which they proposed a query-specific co-ranking framework which can integrate an author-coauthor relationship network and the paper citation network. Compared to their work, our multi-type network provides a more comprehensive framework, and our proposed citation network framework is a global, query-independent one.

PopRank [12] is another representative work whose main idea has been implemented in Microsoft Research Asia Libra [26], a free academic search engine. One advantage of PopRank is that it integrates conferences and journals in addition to authors and papers into consideration. Our framework integrates one additional factor: author affiliation and we combine together content-based analysis and link structure analysis in our framework.

One distinguished contribution of our work, compared to all others discussed above, is that we introduce topical link analysis into consideration. In web research domain, many improvements to PageRank have been proposed, including Topic-Sensitive PageRank [23] in which a separate PageRank score calculation is performed for each topic. With that influence, Nie et al. [18] proposed a Topical PageRank and Topical HITS model which embed topical information into authority propagation and demonstrated better performance over original PageRank and HITS. Even though there has been research work showing use of topical information in analyzing author's publications content (e.g., [17, 16]), no research work, to the best of our knowledge, has introduced topical information into citation network link analysis. We remedy this situation with our paper.

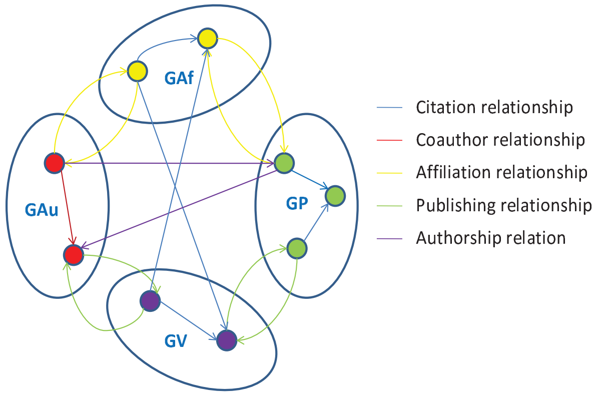

Figure 1: Multi-type (4-T) Citation Network version-1

3. MULTI-TYPE CITATION NETWORK FRAMEWORK

In this section, we introduced the definition of our multi-type citation network framework. Two versions of the framework have been considered, reflecting different relationships among factors.

3.1 Notation and Preliminaries

In the multi-type citation network, different kinds of social factors, as well as their mutual relationships have been considered and integrated. The citation network can be formally denoted as $G = (V, E)$, where

- V is a set of nodes, representing social factors. In our current integrated network, V is combination of four different types of social actors: authors, papers, affiliations and venues.
- E is a set of directed edges, representing relationships among every pair of social actors. All the possible relationships we may have are the relationship between authors, papers, affiliations and venues.

Due to different relationships among the four types of social actors we can consider, we construct two versions of the multi-type citation network, to which we refer as 4-T graph version-1 (4-T) and 4-T graph version-2 (4-TV2) respectively.

3.2 Framework version-1

In 4-T graph version-1, we considered the citation relationship among every pair of social factor types. The graph (shown in Fig. 1) can be viewed as a combination of subgraphs, including those representing each of the types of social factors:

1. Author Graph G_{Au}. There would be one edge from author au_i to author au_j if they coauthored at least one paper or if author au_i cites author au_j. We say that author au_i cites author author au_j if and only if there is at least one publication of au_i that cites one of the publications of au_j. We do not count the number of co-authorship or citationship in the current framework, and thus there would be only one edge between two authors even though they coauthored more than once. The same mechanism works for other subgraphs defined in the following.
2. Paper Graph G_P. There would be one edge from paper p_i to p_j, if p_i cites p_j in its references.
3. Affiliation Graph G_{Af}. There would be one edge from af_i to af_j if two authors, each of which comes from af_i and af_j respectively, coauthor in at least one paper, or there is at least one paper produced in affiliation af_i that cites one of the publications from af_j.
4. Venue Graph G_V. One edge will be drawn from v_i to v_j if there is at least one paper which is published in v_i that cites one of the papers published in v_j.

as well as graphs that relate one type of social actor to another:

1. Bipartite AuthorPaper Graph G_{AuP}. There would be one edge from au_i to p_j, if au_i is one of the authors of p_j. Correspondingly, there would one edge from p_j to au_i, indicating that it is written by au_i.
2. Bipartite AuthorAffiliation Graph G_{AuAf}. One edge would be drawn from au_i to af_j and af_j to au_i, if au_i belongs to the affiliation of af_i. One distinct author may belong to different affiliations in different period of time; thus it is possible for one author node to point to several affiliation nodes.
3. Bipartite AuthorVenue Graph G_{AuV}. If there is at least one paper written by au_i and published in v_j, there would be a corresponding edge from au_i to v_j and from v_j to au_i.
4. Bipartite PaperAffiliation Graph G_{PAf}. One edge will go from paper p_i to affiliation af_j if p_i is written by an author that belongs to af_j.
5. Bipartite PaperVenue Graph G_{PV}. One edge will go from p_i to v_j and v_j to p_i if p_i is published in v_j.
6. Bipartite AffiliationVenue Graph G_{AfV}. If there is one paper belonging to affiliation af_i published in v_j, there would be an edge from p_i to v_j and from v_j to p_i.

3.3 Framework version-2

There may exist redundant information within edges in version-1, since all relationships are generally inferred from the citations among papers. As a result, we introduce a simplified version of the graph.

In this simplified version, we only considered the coauthor relationship between authors, while ignoring the citation relationship between them. Affiliation nodes will only be connected with author nodes, and venue nodes will only be connected with paper nodes. There are no direct edges within the affiliation graph and venue graph. The relationships between authors and venues can be related by firstly relating authors to papers, and then papers to venues. A similar process works when representing the relationship between affiliations and papers. Figure 2 illustrates the simplified version of the multi-type graph.

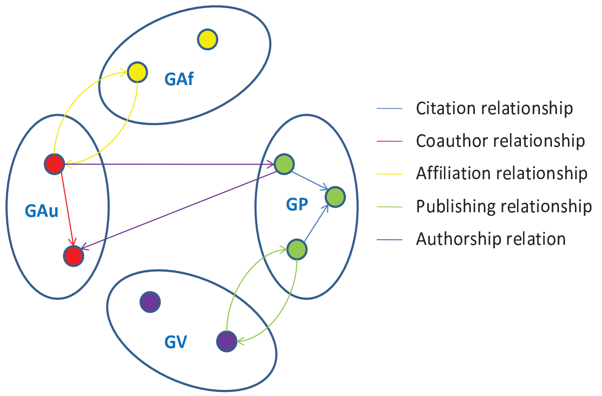

Figure 2: Multi-type (4-TV2) Citation Network

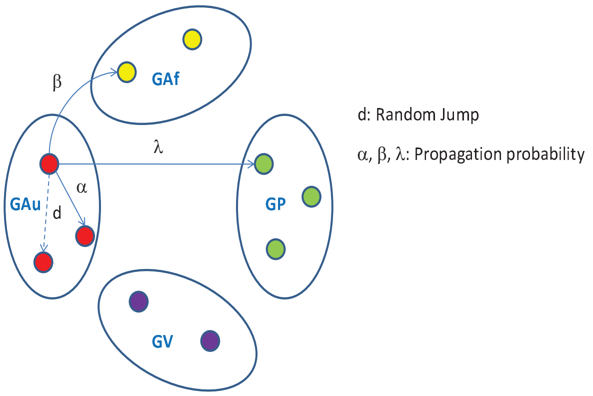

Figure 3: Heterogeneous PageRank

3.4 Heterogeneous PageRank

In the original homogeneous PageRank, each node evenly distributes its authority score among its children. Using such an even propagation in the multi-type citation network, author nodes will evenly distribute its authority to other authors, papers, affiliations, and venues (under framework version-1), which may not well represent the actual interaction possibilities among nodes of different entities. In order to better represent the different impact among multiple types of social actors, we propose a heterogeneous PageRank algorithm based on the assumption that where there would be a different propagation probability for a node to follow different kinds of out-going links (links to different types of nodes). (See Figure 3.) This heterogeneous PageRank can be described as:

$$PR(i) = (1-d)\sum_{j:j\to i}\beta_{ji}\frac{PR(j)}{O(j)_{type(i)}} + d\frac{1}{N} \qquad (1)$$

where:

- j and i are two nodes of any types, where j has outgoing link to i.
- d: random jump.
- β_{ji}: is the parameter determining the propagation probability from node j to i. β_{ji} is equal to β_{jk} if node i and node k are of the same type. $\sum_{type(i)}\beta_{ji} = 1$, where node j has an out-going link to i.
- $O(j)_{type(i)}$ is the number of outlinks j has to the nodes of the same type with i.
- N: total number of nodes in the network.

4. TOPICAL LINK ANALYSIS IN CITATION NETWORKS

4.1 Topical PageRank

The basic idea of Topical PageRank is to incorporate a topic distribution into the representation of each web page as well as the importance score of each page. Therefore, there are at least two vectors associated with each page: the content vector $C_u : [C(u_1), C(u_2), ..., C(u_T)]$, which is a probability distribution used to represent the content of u across T topics, and the authority vector, $A_u : [A(u_1), A(u_2), ..., A(u_T)]$, which is used to measure the importance of the page, where $A(u_K)$ is the importance score on topic K.

Topical PageRank is also a random surfer model. On each page, the surfer may either follow the outgoing links of the page with probability $1-d$ or jump to a random page with probability d. When following links, the surfer may either stay on the same topic to maintain topic continuity with probability α ("Follow-Stay") or jump to any topic i on target page with probability $1-\alpha$ ("Follow-Jump"). The probability of jumping to topic i is determined by $C(u_i)$. When jumping to a random page, the surfer is always assumed to jump to a random topic i ("Jump-Jump"). Thus, the surfer's behavior can be modeled by a set of conditional probabilities:

$$\begin{aligned} P(Follow-Stay|v_k) &= (1-d)\alpha \\ P(Follow-Jump|v_k) &= (1-d)(1-\alpha) \\ P(Jump-Jump|v_k) &= d \end{aligned} \qquad (2)$$

And the probability to arrive at topic i in target page u by the above actions can be described as:

$$\begin{aligned} P(u_i|v_i, Follow-Stay) &= \frac{1}{O(v)} \\ P(u_i|v_k, Follow-Jump) &= \frac{1}{O(v)}C(v_i) \\ P(u_i|v_k, Jump-Jump) &= \frac{1}{N}C(v_i) \end{aligned} \qquad (3)$$

where $O(v)$ represents the out-degree of page v. Therefore, the authority score $A(i)$ on page u is calculated as follows:

$$\begin{aligned} A(u_i) = (1-d)&\sum_{v:v\to u}\frac{\alpha A(v_i) + (1-\alpha)C(v_i)A(v)}{O(v)} \\ &+ \frac{d}{N}C(u_i) \end{aligned} \qquad (4)$$

where $A(v) = \sum A(v_i)$. Note that authors in [18] also proposed a Topical version of the HITS algorithm, which we leave for future work.

4.2 Topical Citation Analysis

Inspired by the principal idea and demonstrated success of Topical PageRank in ranking web pages, we want to introduce such a topical link analysis approach into authors'

Table 1: Queries

algorithms and theory	security and privacy
hardware and architecture	software engineering and programming language
artificial intelligence	machine learning and pattern recognition
data mining	information retrieval
natural language and speech	graphics
computer vision	human computer interaction
multimedia	networks and communications
world wide web	distributed and parallel computing
operating systems	databases
real time and embedded systems	simulation
bioinformatics and computational biology	scientific computing
	computer education

reputation ranking. Similar to web pages, publication papers may also cover different topics, and thus when paper a cites paper b, it may because paper a finds one specific topic t in paper b to be interesting and useful. The same is true for authors' authority propagation. Believing in the prestige of a person on one aspect (say, for example, on data mining) does not mean that this person also owns a high reputation on other aspects (e.g., networking). When authors choose to collaborate and coauthor with each other, they may have mutual trust and interests on some certain aspect (topic). Publishing venues are normally more focused on certain research areas than others. SIGIR, for example, has a high prestige in the information retrieval research field, while SIGCOMM is well-established in the networking domain. Compared to papers, authors or venues, affiliations have less obvious topic-specific differentiation; however, we can still imagine that one affiliation is better known for doing certain kinds of research than others.

5. EXPERIMENTAL WORK

5.1 Data Collection

We used information about papers in the ACM digital library [27] as our experimental dataset and crawled one descriptive web page for each published paper. There are 172,891 distinct web pages within the crawled dataset that appear to have both title and abstract information which we used as our dataset.

For each publication, we extracted and recorded the information of its publishing venue, authors, affiliation of each author, and citation references. Due to possible name ambiguity, we used exact name match to merge author names and conferences, and Jaccard similarity to match affiliations. We extracted 191,386 distinct authors, 45,965 affiliations and 2,197 venues.

In extracting citation references, the title is the representative of each paper, and we only considered those cited papers for which we also crawled the corresponding web page for it. After extracting these factors (paper, authors, affiliations, venues, and the citation relationship among papers), we built two versions of the multi-type citation network as we introduced in Section 3.

5.2 Evaluation

In the portal website of Microsoft Research Asia Libra [26], which is a free computer science bibliography search engine, we found 23 categories (listed in Table 1) covering the main 23 disciplines of computer science research. We used these 23 categories as testing queries, as they represent reasonable topics on which searchers might look for papers, authors, or conferences.

While the link-based citation network analysis is our research focus, we did not use it exclusively for retrieval. Instead, we combined it with the use of a content-based approach. For each author, a profile will be constructed by concatenating all his publications in terms of title, abstract and ACM categories. The Okapi BM25 [8] weighting functions is used to evaluate the relevance between authors' profile with the queries. As a result, for each author, there would be two ranking results: one from using BM25, and the other from a link-analysis approach. These two rankings can then be combined as follows:

$$\lambda * rank_{BM25}(a) + (1 - \lambda) * rank_{CitationNetwork}(a) \quad (5)$$

Since we lack a standard evaluation benchmark for the dataset, we developed three different approaches to consider the relevance between query and search results.

In the first approach, we collected all the PC members in the related conferences for each research area during 2008 and 2009. Libra provides a ranked list of conferences for each of its 23 categories. We retrieved the top 10 conferences for each category and collected their 2008 and 2009 PC members. For those conferences which have no 2008 or 2009 conference, we simply collected the PC members of its two most recently held conferences. To our belief, to be qualified to participate as a PC member is a reasonable indication of the academic reputation of a researcher. To assign different "relevance" scores for those PC members, we normalized across the number of years (two at the most) and the number of different conferences one performs as a PC member.

In the second approach, we collected all the ACM fellows, ACM distinguished and senior members provided from the ACM website. Since there is not research area description for ACM distinguished and senior members, we manually labelled the members into different categories and thus we only used a subset of ACM distinguished and senior members to generate our relevant lists. The subset we retrieved is determined by the mixed group of top 60 ranked authors from all ranking algorithms. Since we focus on top-ranked results in our evaluation metrics, we believe this subset can provide us enough evidence to judge authors. To differentiate the relevance score, we gave a relevance score of 3 to ACM fellows, 2 to distinguished ACM members and 1 for ACM senior members.

We utilized human judgements to generate relevant lists in the third approach. In our evaluation system, the top ten and twenty returned authors by various ranking algorithms were retrieved and mixed together. We then manually but blindly judged the relevance for each author in the mixed ranking list with the corresponding query. Four judges were asked to search using Google Scholar (or other web search engines) using the author name as query, and go through returned webpages (homepages in most cases) related to the author to make a judgment on his professional prestige.

After generating the relevant lists, we can compute and compare the retrieval and ranking performance of different ranking algorithms. We took the well-known metric, the Normalized Discounted Cumulated Gain (NDCG) as our

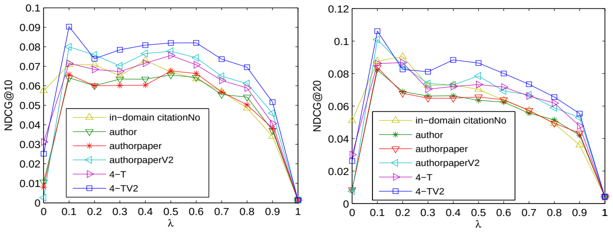

Figure 4: Comparison among different levels of citation network (NDCG@10 and NDCG@20 for ACM members) as the BM25 weight (λ) is varied.

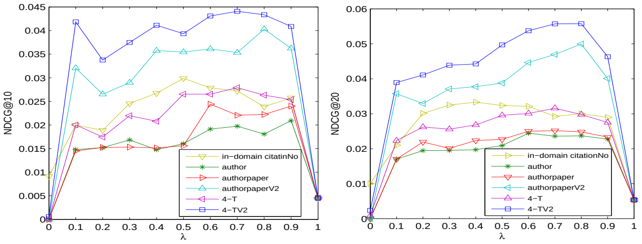

Figure 5: Comparison among different levels of citation network (NDCG@10 and NDCG@20 for PC members) as the BM25 weight (λ) is varied.

main metric. We tested on NDCG@10 and NDCG@20 respectively.

5.3 Experimental Results

We made several groups of comparisons to test the performance of different algorithms in their abilities of finding the most influential authors in 23 different research fields (represented as queries).

5.3.1 *Multi-type Citation Network*

Figures 4-5 indicate the NDCG results for different kinds of citation network analysis approaches using original uniform PageRank as propagation mechanism and using ACM members and PC members as evaluation dataset respectively. Table 2 shows the results from human judgements. To reduce the amount of manual labelling, we only gave to judges the results when combined with BM25 with parameter λ set to 0.5. We also introduced the ranking method of in-domain citation count as one of our compared approach. We took the 23 categories provided by Libra as domains, and regarded it as in-domain citation if two papers are within one domain and there is a citationship between them.

Table 2: NDCG Results from human judgements (λ=0.5)

citation graph	NDCG@10	NDCG@20
in-domain citationNo	0.6820	0.6748
author	0.6390	0.6025
atuhorpaper	0.6455	0.6167
authorpaperV2	0.6899	0.6614
4-T	0.6545	0.6401
4-TV2	0.6988	0.6889
Topical 4-T	0.7004	0.6848
Topical 4-TV2	0.7490	0.7231

Table 3: Top-level topics from the ACM Digital Library.

computer applications
computer aided engineering
computing milieux
general literature
information systems
software
computer systems organization
computing methodologies
data
hardware
mathematics of computing
theory of computation

Several conclusions can be drawn from these results. First, there is a noticeable consistency with regard to the performance of ranking algorithms for the three different evaluation methods. 4-TV2 always works the best in all scenarios. This demonstrates our initial intuition that affiliations and conference venues can provide useful information and thus make them important and unignorable social factors in determining authors' reputations, and that the mutual reinforcement among different factors can improve ranking performance.

Secondly, we also noticed that different versions of the citation graph do have different impact on the overall performance. From the above figures, we find that version2 always work better than version1. This may be caused by the fact that removal of possible duplicate citation relationships can avoid authority being scattered over duplicate links. The results give us an indication that we should not only consider increasing the number of social factors to explore, but also need to pay attentin to how to effectively find the relationships among them and thus properly organize them.

We noticed that the absolute NDCG values for ACM members and PC members are comparatively low, but this may be caused by the incomplete collection of papers from the ACM digital library, and the incomplete citation relationships we collected. As we mentioned before, we only took those citations for which we have also crawled the corresponding web page into account. Besides, some distinguished researchers may have published in many journals or other papers which are not normally collected by the ACM digital library. However, since there is a high consistency among all the different evaluation approaches, and the NDCG value of using human judgement is pretty high, we can have confidence in the evaluation using ACM members and PC members.

5.3.2 *Topical PageRank*

A key issue in Topical PageRank is to generate the static per-document content vector. We made use of the hierarchically-organized ACM categories provided by authors of each paper for topic distribution generation. We extracted the top level primary category and additional category for all the papers in the dataset and thus get 12 categories in total (listed in Table 3). We regard these categories as topics. With category information provided, computing topic distributions for papers is straightforward.

Since each author is represented by a profile which is a concatenation of all the papers he has written, we can accumulate all the topics mentioned by each published paper, and then compute the topic distribution. The same mechanism works for computing the topic distribution of venues, for which we collected all the papers published in that venue, accumulated papers' topics, and then computed the corresponding distribution. We did the same thing for generating affiliations' topic distribution by collecting the papers writ-

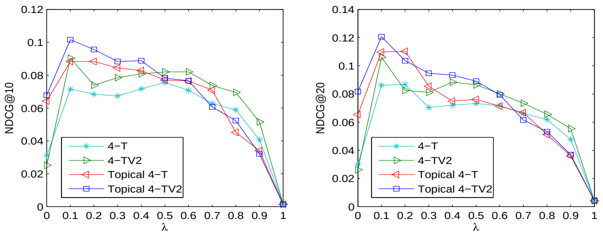

Figure 6: Topical PageRank Performance (NDCG@10 and NDCG@20 for ACM members) as the BM25 weight (λ) is varied.

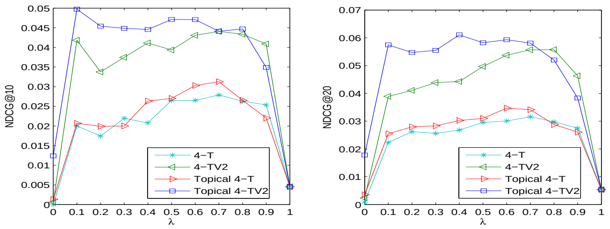

Figure 7: Topical PageRank Performance (NDCG@10 and NDCG@20 for PC members) as the BM25 weight (λ) is varied.

ten by authors from that affiliation, and taking use of the papers' topic distribution to compute the affiliations' topic distribution.

Since we take the categories provided by Libra as experimental queries, and they also list group of papers for each category, we randomly chose five papers from each category and take use of papers' topic distribution to compute the topic distribution for queries.

See Figures 6-7 for the results of topical experiments.

We set the α value to be 0.85 in all experiments. We found once again a high consistency among the results from different approaches , and that introducing Topical PageRank can improve the performance indeed. The improvement of the best performance of Topical 4-TV2 over that of 4-TV2 is 12.9% (NDCG@10) and 14.2% (NDCG@20) for ACM members, 12.7% (NDCG@10) and 9.7% (NDCG@20) for PC members, and 6.8% (NDCG@10) and 5.1% for human labelling results.

5.3.3 *Comparison with two baselines*

One of the main characteristics of our multi-type citation network analysis approach lies in its combination of both content-based approach and graph-based approach. We took two other approaches as our comparison baselines, one is BM25, a purely content-based approach, and the other is the CoRank approach proposed in [25].

The results of incorporating BM25 has been shown in all the above figures, since it is equivalent to pure BM25 when λ is set to be 1. As we can see, our multi-type citation network outperforms BM25 in all different scenarios.

The CoRank algorithm generates author and paper rankings by taking propagation between authors and papers into account. It is a graph-based approach. Instead of building a big graph for all the authors and papers in the dataset, they first rank authors in terms of their topic weights in a certain domain, retrieve the top 500 authors, and build the graph based on these authors and their publications. The graph they build is thus query-specific. We have implemented this algorithm (we determined the topic weight by counting the number of papers belonging to a topic (query)), and Table 4 shows the comparion results between CoRank and our TopicalV2 at its best performance. As we can see, TopicalV2 outperforms that of CoRank. We used PC members for evalution in this experiment.

Table 4: TopicalV2 vs CoRank (on PC members)

Citation graph	NDCG@10	NDCG@20
Topical 4-TV2	0.0497	0.0611
CoRank	0.0219	0.0308

Table 5: NDCG@20 for Heterogenous PageRank

Best Perf.	Parameter settings							Perf.
on Training	p1	p2	p3	p4	p5	p6	p7	on Test
0.0944	0.4	0.1	0.5	0.2	0.3	0.5	0.6	0.0905
0.0868	0.6	0.1	0.3	0.4	0.1	0.5	0.5	0.1262
0.1108	0.6	0.1	0.3	0.4	0.1	0.5	0.4	0.0040
0.0957	0.4	0.3	0.3	0.4	0.1	0.5	0.2	0.0929
0.1045	0.6	0.1	0.3	0.4	0.1	0.5	0.4	0.0465
Average performance on Test								0.0720

5.3.4 *Heterogeneous PageRank*

We proposed a heterogenous PageRank algorithm with the intention of exploring on the different impact among social factors. The parameter β_{ij} indicates the authority propagation probability among factors i and j. It is actually a parameter optimization problem if we want to get the best performance by tuning the parameters.

We worked on 4-TV2, and thus there are totally seven parameters, the propagation probability between authors to authors (p1), authors to papers (p2), authors to affiliations (p3), papers to authors (p4), papers to papers (p5), papers to venues (p6), and the combination parameter with BM25 (p7). We performed greedy search, testing on the possible combinations of the parameters with a stepsize of 0.2 (the combination parameter p7 with BM25 has a stepsize of 0.1). In order to test on the system performance on unseen data, we further group the 23 queries into 5 groups, and used five-fold cross validation approach to evaluate system performance. We evaluated on PC member-based evaluation.

The algorithm under different parameter scenarios converges within 8-17 iterations. As indicated in Table 5, the average performance using heterogeous PageRank is even better than the best performance of topical 4-TV2 (0.0611, which is the best performnace in all the previous experiments). This improvement is around 17.8%. This demonstrates our initial intuition that considering different effect among factors can improve performance.

6. CONCLUSIONS

Previous work has investigated the value of integrating author and paper information in citation network in ranking authors' reputations. PopRank is a work we have identified which integrated conference venues into consideration. We further observed that there are yet more useful information we can extract and take use of, for example, affiliations. To test this idea, we proposed in this paper a multi-type citation network framework which integrates citations among authors, papers, affiliations and publishing venues into one model, and uses a PageRank-based algorithm to rank authors' authority. In order to test the different impact among factors, we further proposed a heterogeneous PageRank algorithm in which social factors may propagate authority to neighboring factors with different probabilities. Moreover, in order to better evaluate the prestige of an author in different kinds of research topics, we incorporated topical link analysis into the citation network. We conclude from experimental results that:

- Multi-type citation networks can effectively improve ranking performance. Affiliation and publishing venues provide additional useful information in evaluating authors' reputations.
- Topical link analysis shows positive improvement in ranking authors' authority.
- Heterogeneous PageRank, with parameter tuning, can work even better than Topical PageRank.

Despite the current improvements shown, there are many possible future directions, including:

- A finer citation framework and link analysis mechanism. For example, we can assign different weights to edges connecting authors to authors, to represent the number of times they coauthored.
- A different ranking algorithm. HITS may be a good choice, since it well models the mutual reinforcement from two communities.
- A better mechanism could be used to disambiguate names.
- Moreover, other topical models can be adapted.

Acknowledgments

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

7. REFERENCES

[1] B. D. Davison. Toward a unification of text and link analysis. In *Proc. ACM SIGIR*, pages 367–368, 2003.

[2] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(60):471–479, November 1972.

[3] E. Garfield. Citation indexing: Its theory and application in science, technology, and humanities. John Wiley and Sons, Inc., New York, NY, USA, 1979.

[4] P. Glenisson, W. Gldnzel, F. Janssens and B. De Moor. Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing and Management: an Int'l Journal*, 41(6):1548-1572, Dec. 2005

[5] G. Pinski and F. Narin. Citation influence for journal aggregates of scientific publications: Theory with applications to literature of physics. *Inf. Proc. and Man.*, 1976.

[6] J. E. Hirsch. Citation indexing: Its theory and application in science, technology, and humanities. In *Proc. of the National Academy of Sciences*. John Wiley and Sons, Inc., New York, NY, USA, 2005.

[7] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of ACM*, 46(5):604–632, 1999.

[8] S. Robertson. Overview of the OKAPI projects. *Journal of Documentation*, 53:3-7, 1997.

[9] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proc. of the 29th Annual Int'l ACM SIGIR Conf. on Research and Development on Information Retrieval*, pages 43–50, 2006.

[10] H. Fang and C. Zhai. Probabilistic models for expert finding. In *ECIR*, pages 418-430, 2007.

[11] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *Proc. of the 15th ACM Int'l Conf. on Information and knowledge management*, pages 387-396, 2006.

[12] Z. Nie, Y. Zhang, J. Wen, and W. Ma. Object-Level Ranking: Bringing Order to Web Objects. In *Proc. of the 14th Int'l World Wide Web Conf. WWW*, pages 567-574.

[13] W. Xi, B. Zhang, Y .Lu, Z. Chen, S. Yan, H .Zeng, W. Ma, and E. Fox. Link Fusion: A unified link analysis framework for multi-type interrelated data objects. In *Proc. of the 13th Int'l World Wide Web Conf. WWW*, 2004.

[14] Z. Guan, J. Bu, Q. Mei, C. Chen, and C. Wang. Personalized Tag Recommendation Using Graph-based Ranking on Multi-type Interrelated Objects. *Proc. of the 32nd Annual Int'l on Research and Development in Information Retrieval SIGIR*, pages 540-547.

[15] X. Liu, J. Bollen, M. L. Nelson, and H. V. de Sompel. Coauthorship networks in the digital library research community. In *arXiv.org:cs/0502056*, 2005.

[16] G. S. Mann, D. Mimno, and A. McCallum. Bibliometric impact measures leveraging topic analysis. In *Proc. JCDL*, 2006.

[17] D. Mimno and A. McCallum. Mining a digital library for influential authors. In *Proc. JCDL*, 2007.

[18] L. Nie, B. D. Davison, and X. Qi. Topical link analysis for web search. In *Proc. of the 29th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 91–98, 2006.

[19] P.Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with Google's PageRank algorithm. In *Journal of Informetrics*, 2007.

[20] 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*, November 1998.

[21] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*. Cambridge University Press, 1994.

[22] T. Bogers, K. Kox, and A. van den Bosch. Using citation analysis for finding experts in workgroups. In *DIR2008*, 2008.

[23] T. H. Haveliwala. Topic-sensitive PageRank. In *in WWW'02: Proc. of the 11th Int'l Conf. on World Wide Web*, pages 517–526. ACM, 2002.

[24] X. Wang, J. Sun, Z. Chen, and C. Zhai. Latent semantic analysis for multiple-type interrelated data objects. In *Proc. ACM SIGIR*, pages 236–243. ACM, 2006.

[25] D. Zhou, S. A. Orshanskiy, H. Zha, and C. Giles. Co-ranking authors and documents in a heterogeneous network. In *IEEE Int'l Conf. on Data Mining (ICDM)*, 2007.

[26] Libra. Now known as Microsoft Academic Search, 2009. Available via http://libra.msra.cn/ and http://academic.research.microsoft.com.

[27] ACM Digital Library, 2009. http://portal.acm.org/dl.cfm.