ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st&22ndMarch Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

# Co-Occurrence Diffusion Method for Expert Ranking Algorithm in Heterogeneous Web Pages

Dr. VijayaChamundeeswari[1] , Ramyadevi.R[2] , Rajkumar.S.C[3]

Faculty[1], Department of CSE, Velammal Engineering College, Anna University, Chennai, India
Student[2], Department of CSE, Velammal Engineering College, Anna University, Chennai, India

**ABSTRACT:-**Web-based communities have become important places for people to searching experts on the web, Page Rank algorithm has proven to be very effective for ranking Web pages, and the rank scores of Web pages can be manipulated. The Web spam problem which refers to hyperlinked pages on the Web that are created with the intention of misleading search engines. The reason for the increasing amount of Web spam is explained in some web site operators try to influence the positioning of their pages within search results because of the large fraction of web traffic originating from searches and the high potential monetary value of this traffic including PageRank on this large size social network in order to identify users with high. To handle the manipulation problem and to cast a new insight on the Web structure, we propose of utilizing co occurrence relationships to assess the relevance and reputation of a person name with respect to a query on document simultaneously search based on the ranking algorithm called Diffusion Rank ,it is motivated by the heat diffusion phenomena, which can be connected to Web ranking because the activities flow on the Web can be imagined as heat flow, we generate a ranked list of people's name and leave the person identification problem to users.

**Index Terms** - Expert search, web mining, Page Rank, Diffusion Rank, hyper graph,Expertise ranking, probabilistic model, heterogeneous bibliographic network

## I.INTRODUCTION

Finding Experts on the web many approaches have been proposed and shown to be useful for expertise ranking .The most popular and successful approaches obtain their estimator of personal expertise by aggregating the relevance scores of documents directly associated with a person, which could be classified as document-centric models [6]. These methods share the underlying claiming that the relevance of the textual context of a person adds up to the evidence of his/her expertness, but they only consider the textual documents while ignoring the relations between experts as well as valuable heterogeneous networks. For example, Searching the Web for names of people can be a challenging task when a single name is shared by many people. The task of finding people who are experts on a topic has recently received increased attention. Introduce a different expert finding task for which a small number of example experts is given (instead of a natural language query) and the system's task is to return similar experts. It show that more fine grained representations of candidates result in higher performance, and larger sample sets as input lead to improved precision while the Page Rank algorithm [1] has proven to be very effective for ranking Web pages, inaccurate Page Rank results are induced because of web page manipulations by people for commercial interests. The manipulation problem is also called the Web spam From the viewpoint of the Web site operators who want to increase the ranking value of a particular page for search engines, Keyword Stuffing and Link Stuffing are being used widely [1,2]. There are two methods being employed to combat the Web spam problem. Machine learning methods are employed to handle the keyword stuffing. To successfully apply machine learning methods, to mark part of the Web pages as either spam or non-spam, then to apply supervised learning techniques to mark other pages. Link analysis methods are also employed to handle the link stuffing problem.To concern about brief literature review on various related ranking techniques.Establish the Heat Diffusion Model (HDM) on various cases in Section 3, propose Diffusion Rank in Section 4. In Section5 conclusions

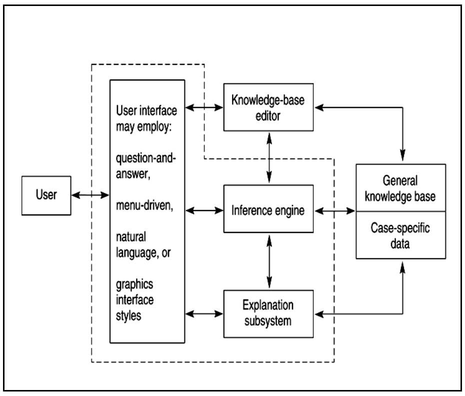


**Fig 1. General Expert Search**

## Co-Ranking Authors and Documents in a Heterogeneous Network

A framework for co-ranking entities of different kinds in a heterogeneous network connecting the researchers (authors) and publications they produce (documents). The heterogeneous network is comprised of GA, a social network connecting authors, GD, the citation network connecting documents, and GAD, the bipartite authorship network

**(i)First Model** is equivalent to a profile-centric approach where text from all the documents associated with a person is amassed to represent that person.

**(ii)Second Model** is a document-centric approach which first computes the relevance of documents to a query and then accumulates process was formulated in a generative probabilistic model for each person the relevance scores of the documents that are associated with the person.

Researchers have investigated using additional information to boost retrieval performance, such as PageRank, indegree, and URL length of documents [1,2], person-person similarity [1], internal document structures that indicate people’s association with document content [1], query expansion and relevance feedback using people names, nonlocal evidence [2], [7], proximity between occurrences of query words and people names [19], [3].Besides language models, other methods have been proposed for organizational expert retrieval between people and documents as a bipartite graph

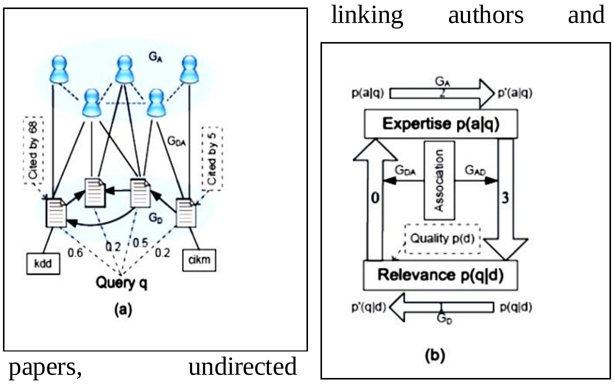


**Fig 2(a) Heterogeneous [7] (b)Basic network models[7]**

However, the characteristics of the heterogeneous networks as shown in Figure 1(a) are not fully exploited by these methods. On one hand, the link structures among persons and documents are different which should be treated differently Figure 1(a) presents an example heterogeneous network according to a scientific bibliography, with undirected authorship graph GDA linking authors and papers, undirected publishing edges linking papers and venues, and directed graph GD linking papers to other papers, as well as co-authorship graph GA linking authors and other authors. The linking edges among these components reveal a lot of valuable information about the potential relevance and expertise propagation over the network. Our idea is that different types of edges provide different information which should be treated differently. A basic framework of our model is illustrated in Figure 1(b). First, the citation graph GD can be used to refine the relevance in the document level, then the relevance of documents of a person adds up to the evidence of his/her expertise based on the undirected

Authorship graph GDA. Second, the co-authorship

Graph GA can be used to refine the expertise in the person level. Third, it is essential to check whether the expertise of persons can reinforce the relevance of their associated documents with respect to a query.

The contributions of this paper include:

(1) A new framework for co-ranking entities of two types in a heterogeneous network is introduced

(2) The framework is adapted to ranking authors and documents: a more flexible definition of the social network connecting authors is used and random walks that are part of the framework are appropriately designed for this particular application (3) Empirical evaluations have been performed on a part of the Cite-Seer data set allowing to compare co-ranking with several existing metrics. Obtained results suggest that co-ranking is successful in grasping the mutually reinforcing relationship,

Therefore making the rankings of authors and documents depend on each other.

## II. RELATED WORK

The importance of a Web page is determined by either the textual content of pages or the hyperlink structure or both. As in previous work [1, 6], we focus on ranking methods solely determined by hyperlink structure of the Web graph. All the mentioned ranking algorithms are established on a graph. For our convenience, we first give some notations. Denote a static graph by G = (V,E), where V ={v1,v2 ……. Vn} E = {(vi, vj)| there is an edge from vi to vj}. Ii and di denote the in-degree and the out-degree of page i respectively.

### 1. Page Rank

A method is used for rating the web pages importance of objectively and mechanically using the link structure of the web. When searching information on the Web, An user query is an input for the search engine. The search engine is return as the query's result, a list of Web sites which usually is a huge set. So the ranking of these web sites is very important. Because much information is contained in the link-structure of the Web, information such as which pages are linked to others can be used to augment search algorithms.

**Page Rank Search Engines:**

**Title Search Engine** Searches only the "Titles" Finds all the web pages whose titles contain all the query words Sorts the results by PageRank Sorts the results by PageRank Very simple and cheap to implement Title match ensures high precision, and PageRank ensures high quality

**Full Text Search Engine** It examines all the words in every stored document and also performs PageRank (Rank Merging) More precise but more complicated is Called Google. Every page has some number of Forward links (outedges) and Back links (inedges)

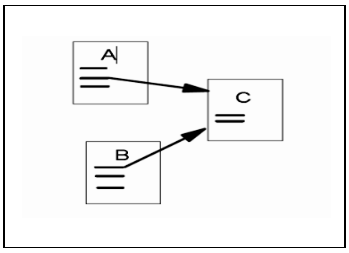


**Fig 3. Forward and Backward links**

A and B are C's backlinks & C is A and B's forward links.

Generally, highly linked pages are more "important" than pages with few links. For example, if a web page has a link off the yahoo home page, it may be just one link but it is a very important one. A page has high rank if the sum of the ranks of its backlinks is high. This covers both the case when a page has many backlinks and when a page has a few highly ranked backlinks[1].

### 2. Trust Rank

Trust Rank [2] is composed of two parts. The first part is the seed selection algorithm, in which the inverse Page Rank was proposed to help an expert of determining a good node. The second part is to utilize the biased Page Rank, in which the stochastic distribution g is set to be shared by all the trusted pages found in the first part. Moreover, the initial input of x is also set to be g. The justification for the inverse Page Rank and the solid experiments support its advantage in combating the Web spam. Although there are many variations of Page Rank, e.g., a family of link-based ranking algorithms in [2], Trust Rank is especially chosen for comparisons for three reasons: (1) It is designed for combating spamming (2) It fixed parameters make a comparison easy and (3) It has strong theoretical relations with Page Rank and Diffusion Rank.

### 2. Hits Algorithm

Hypertext Induced Topic Search (HITS) or hubs and authorities is a link analysis algorithm to rate Web pages.[8] A precursor to PageRank, HITS is a search query dependent algorithm that ranks the web page by processing its entire in links and out links . Thus, ranking of the web page is

(i)Authority: pages that provide important, trustworthy information on a given topic

(ii)Hub: pages that contain links to authorities

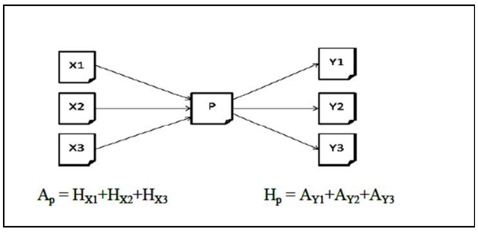


**Fig 4. Authority And Hub**

1. Start with each node having a hub score and authority score of 1.
2. Run the Authority Update Rule
3. Run the Hub Update Rule
4. Normalize the values by dividing each Hub score by the sum of the squares of all Hub scores, and dividing each Authority score by the sum of the squares of all Authority scores.
5. Repeat from the second step as necessary.

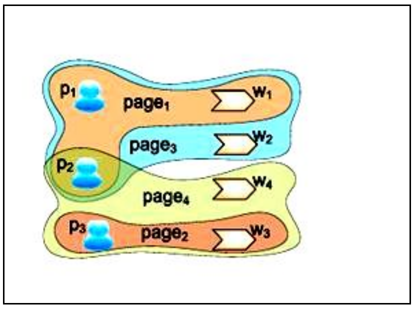
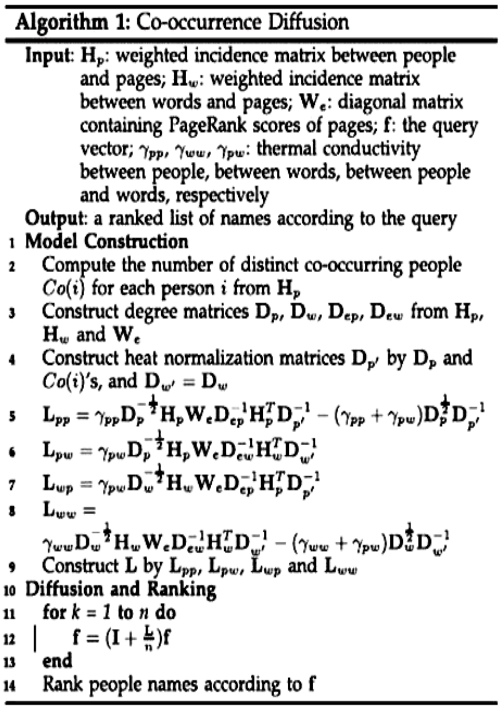


**Fig 5.Heterogeneous example**

given P, W, $G_{P,W}$ and query keywords from W, to rank P according to their expertise in the topic represented by the query

### 3. Diffusion Rank

**Algorithm**Diffusion Rank Function Input: The transition matrix A the inverse transition matrix U the decay factor $\alpha_I$ for the inverse Page Rank the decay factor $\alpha_B$ for Page Rank number of iterations MI for the inverse Page Rank the number of trusted pages L, the thermal conductivity coefficient γ Output: Diffusion Rank score vector h.For the ranking task, we adopt the heat kernel on a random graph. Formally the Diffusion Rank is, in which, the element $U_{ij}$ in the inverse transition matrix U is defined to be $1/I_j$ if there is a link from i to j, and 0 otherwise.



```
Algorithm 1: Co-occurrence Diffusion
Input: $H_p$: weighted incidence matrix between people
       and pages; $H_w$: weighted incidence matrix
       between words and pages; $W_e$: diagonal matrix
       containing PageRank scores of pages; f: the query
       vector; $\gamma_{pp}, \gamma_{ww}, \gamma_{pw}$: thermal conductivity
       between people, between words, between people
       and words, respectively
Output: a ranked list of names according to the query
Model Construction
    Compute the number of distinct co-occurring people
    $Co(i)$ for each person $i$ from $H_p$
    Construct degree matrices $D_p, D_w, D_{ep}, D_{ew}$ from $H_p$,
    $H_w$ and $W_e$
    Construct heat normalization matrices $D_{p'}$ by $D_p$ and
    $Co(i)$'s, and $D_{w'} = D_w$
    $L_{pp} = \gamma_{pp} D_p^{-\frac{1}{2}} H_p W_e D_{ep}^{-1} H_p^T D_{p'}^{-1} - (\gamma_{pp} + \gamma_{pw}) D_p^{\frac{1}{2}} D_{p'}^{-1}$
    $L_{pw} = \gamma_{pw} D_p^{-\frac{1}{2}} H_p W_e D_{ew}^{-1} H_w^T D_{w'}^{-1}$
    $L_{wp} = \gamma_{pw} D_w^{-\frac{1}{2}} H_w W_e D_{ep}^{-1} H_p^T D_{p'}^{-1}$
    $L_{ww} =$
    $\gamma_{ww} D_w^{-\frac{1}{2}} H_w W_e D_{ew}^{-1} H_w^T D_{w'}^{-1} - (\gamma_{ww} + \gamma_{pw}) D_w^{\frac{1}{2}} D_{w'}^{-1}$
    Construct L by $L_{pp}, L_{pw}, L_{wp}$ and $L_{ww}$
Diffusion and Ranking
    for $k = 1$ to $n$ do
        $f = (I + \frac{L}{n}) f$
    end
    Rank people names according to f
```

**[7]Co-occurrence Diffusion Algorithm**

## III. EXPERIMENTS

### 1. Data Preparation

Experimental data sets were extracted from the ClueWeb09 web collection which is a result of recent web crawl and consists of about 1.04 billion webpages in 10 languages. but only considered the 500 million English webpages. PageRank scores were computed based on the link graph among all the 500 million English webpages. For people names, we extracted author names from the Digital Bibliography & Library Project (DBLP) bibliography data set.3 The reasons that we use DBLP author names are:

1) It contains a large number of names, 800 K names
2) Both senior and junior people (i.e., experts and no experts)
3) It is easy to construct ground-truth data sets for evaluation. The process for generating our experimental data sets is as follows: first we did a sequential scan through all the 500 million English webpages to extract all the occurrences of author names, where simple rules, "First Middle Last" and "Last, First Middle," are used to find name occurrences. We discarded names which did not appear in those English pages. After this step we got 520,971 distinct people names and 37 million pages, each of which contains at least one person name. Extracted and processed those pages' text content and built index for them. Then we selected, from the remainder, the webpages that contain at least five distinct people names and at least
30 distinct words, in order to reduce data set size. This yields 3,608,265 pages and 478,896 names. Our task is to find top-10 or 20 among these names for a given query. To provide a notion of where these 3,608,265 pages come from, It show the top five domains of those pages in Fig 6. We can see that a large amount of webpages do not come from academic websites (e.g., .edu). Formulated three data sets from these pages:

1) DATA-3M: This contained all 3,608,265 pages
2) DATA-1M: This consisted of a random subset of 1 million pages from the 3,608,265 pages
3) DATA-0.2M: This consisted of a random subset of 200k pages from the 3,608,265 pages. Using DATA-1M for most of the experiments. DATA-3M and DATA-0.2M were used to investigate the influence of data sizes on the performance

### 2. Evaluation Methodology

The first two algorithms are simple heuristics which follow the intuition about topical experts discussed The first one, which is called NameFreq, computes the total number of times a name appears in pages that contain all the query keywords. Frequency in each page is weighted by the corresponding PageRank score. Thus, NameFreq actually computes the for a person name i in a query-dependent local context. The second one, NameCoFreq, counts the number of distinct names which co-occur with a name in pages containing all the query keywords. The third one is the language model-based algorithm proposed in [6], which is one of the most prominent methods for organizational expert search, denoted by LM. computer science. The 24 area names are treated as test queries and the corresponding top 100 authors are taken as ground truth expert lists. The other one is a manually labeled ground-truth data set used in [7], which contains 17 queries and the averaged number of experts for each query is 29.35. Refer to the two benchmark data sets as Libra-GT and Manual-GT, respectively. Libra-GT contains more general queries while Manual-GT contains more specific ones. There are 41 queries in total.

## 3. Author Rankings

To evaluate the co-ranking approach, Perform a ranking of authors in each topic t by the methods listed below:

1)Publication count, the number of papers (on the topic t) an author has in the document subset.

2) Topic weight, the sum of topic weights of all documents, produced or co-authored by an author

3) Number of citations, the total number of citations to the documents of an author from the other documents on the same topic

4) PageRank in the social network, ranking by PageRank on the graph GA, constructed as outlined.

5) Co-Ranking, co-ranking authors and documents by the new method. The parameter values used in the Co-Ranking framework are m = 2, n = 2, k = 1, _ = (λ2, γ= 0.1. For different settings of m, n, k the top 20 authors and papers varied slightly, even less for differentγ. A well-known metric, the Discounted Cumulated Gain (DCG) [6], in order to compare the five different rankings of authors. Top 20 authors according to each ranking (publication count, etc.) are merged in a single list, shuffled and submitted for judgment. Two human judges, one an author of this paper and the other one from outside, provide feedback. Numerical assessment scores of 0, 1, 2, and 3 are collected to reflect the judge's opinion with regard to whether an author is ranked top 20 in a certain field, which respectively means strongly disagree, disagree, agree, and strongly agree, with the fact that these authors are ranked top 20 in the corresponding field. As suggested, assessments were carried out based on professional achievement of the authors such as winning of prestigious awards.

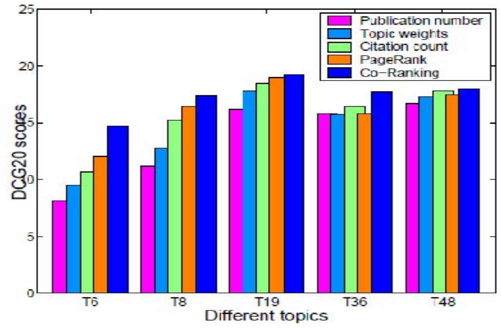


**Fig 6. DCG$_{20}$ scores for author rankings: number of papers, topic weights, number ofcitations, PageRank, and Co-Ranking.**

## IV. Running Time

The running time of Co-Diffusion when varying the number of relevant pages in Local Ranking. The experiment is run on a PC with Intel Core i7 CPU and 12 GB memory. The number of iterations is set to 100, which is sufficient for all the queries in our experiments. Can see the running time of CoDiffusion grows approximately linearly with the number of relevant pages. This is consistent with our complexity analysis in Varying the number of webpages only changes the number Running time when varying the number of relevant webpages of nonzero elements in Hp, Hw, and L. Diffusion and Ranking costs more time than Model Construction. This is because L is not only larger, but also much denserthan Hp and Hw. CoDiffusion cannot outperform the baseline algorithms in terms of running time. The running time of RW is shown in RW is more efficient since its time cost depends on the number of nonzero elements in Hp which is much sparser than L. Discuss the scalability issue shortly.

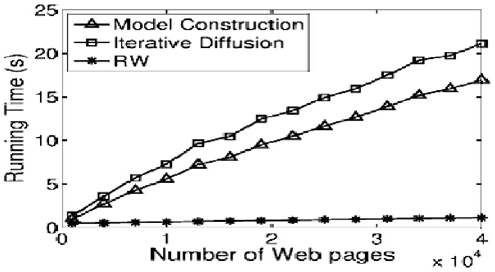


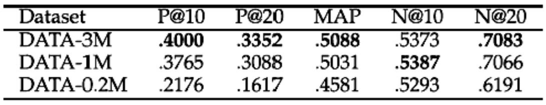
**Fig 7. Runtime Graph**

The running time of Global Ranking on DATA-1M are 190 and 245 s, for Model Construction and Diffusion, respectively. This is because L with more pages is much denser.

## V. RESULT ANALYSIS

Performance Comparison of Co-Diffusion on Different Sizes of Data Sets with Respect to Manual-GT **Table 1**

**Performance Comparison of Co-Diffusion on Different Sizes of Data Sets with Respect to Libra-GT**



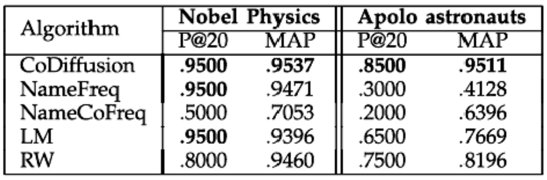
| Dataset | P@10 | P@20 | MAP | N@10 | N@20 |
|---|---|---|---|---|---|
| DATA-3M | **.4000** | **.3352** | **.5088** | .5373 | **.7083** |
| DATA-1M | .3765 | .3088 | .5031 | **.5387** | .7066 |
| DATA-0.2M | .2176 | .1617 | .4581 | .5293 | .6191 |

Clearly, these ranking lists are not as good as the ranking list generated by CoDiffusion, although they all put the most famous swimmer "Michael Phelps" at the top position. In particular, "Mike James" is a popular name and shows up relatively frequently in the related pages. Ana Ivanovic is a former World No. 1 tennis player. However, Can also find many co-occurrences between her name and "Swimming." For example, her Wikipedia page says "she admitted that she trained in an abandoned swimming pool.Although these two names appear frequently in related pages, they do not get into top four of NameCoFreq. "St. Thomas" appears frequently on the web as an university name, although it is an author name in DBLP. "Juan Carlos" is a popular Spanish name and it also refers to different things (e.g., it is a part of a university8 name). Hence, they co-occur with a lot of different names. Nevertheless, they do not get into the top four of NameFreq.

Our algorithm successfully makes use of different kinds of co-occurrence information to return top swimmers in the highest positions. consider two queries "Nobel Physics" and "Apolo astronauts."For "Nobel Physics," obtain names of people who have won Nobel prizes in Physics from http://nobelprize.org/nobel_{p}rizes/physics/laureates/ for "Apolo astronauts" get the list of all Apolo astronauts from Wikipedia

**Table 2**

**Performance Comparison of Expert Search Algorithms on Two Queries: "Nobel Physics" and "Apolo Astronauts"**



| Algorithm | Nobel Physics | | Apolo astronauts | |
|---|---|---|---|---|
| | P@20 | MAP | P@20 | MAP |
| CoDiffusion | **.9500** | **.9537** | **.8500** | **.9511** |
| NameFreq | **.9500** | .9471 | .3000 | .4128 |
| NameCoFreq | .5000 | .7053 | .2000 | .6396 |
| LM | **.9500** | .9396 | .6500 | .7669 |
| RW | .8000 | .9460 | .7500 | .8196 |

These names are treated as the ground truth. Our name set is extended to include all the groundtruth names. Although adding ground-truth names can lead to optimistic ranking results, it is fair for all the algorithms. The experimental results are shown in Table 3. As can be seen, for "Nobel Physics" almost all the algorithms can achieve good performance. NameFreq and LM do as good as CoDiffusion in terms of P@20. However, CoDiffusion has a better MAP, indicating it gives a better ranking. Regarding "Apolo astronauts," CoDiffusion performs much better than Performance Comparison of Expert Search Algorithms on Two Queries: "Nobel Physics" and "Apolo Astronauts"

## VI. DISCUSSIONS

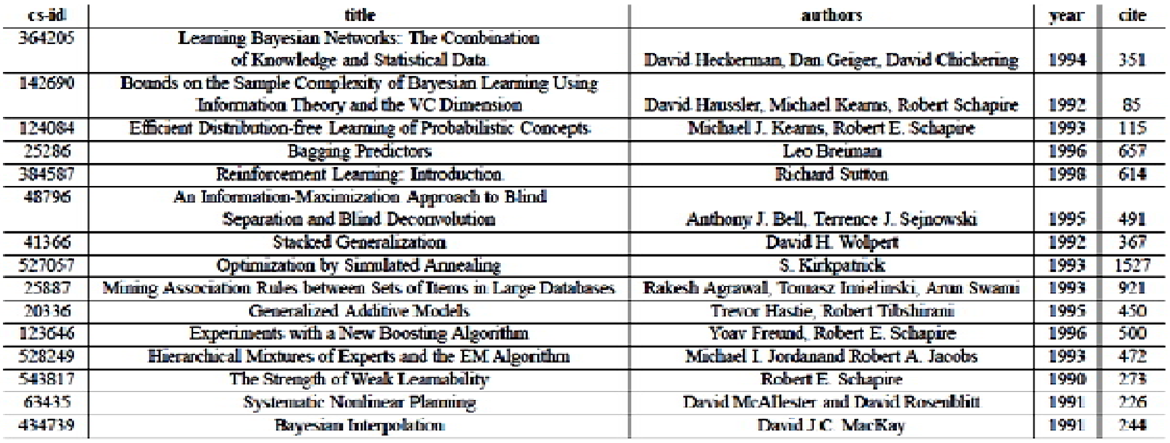
Expert search on the web is intrinsically different from enterprise expert search. ordinary webpages could be noisy and contain vague expertise evidences.

1)Association scores among people and words can be further adjusted by advanced techniques such as NLP through customizing the thermal conductivity for each pair of objects [7]

2)Other page quality measures [6] can also be integrated through the hyperedge weighting scheme. However, not going to explore these possible enhancements in this work. In the enterprise expert search, person identification is not difficult: can obtain e-mail addresses or employee identifiers to uniquely identify an employee. The complete list of employees is known in advance. However, it is difficult to identify people on the web as names are more available than e-mail addresses. In this work, generate a ranked list of people's name and leave the person identification problem to users. With a returned name list, users can identify experts by searching their names together with the query topic through a web search engine. use a set of names extracted from DBLP to bypass the name extraction problem, which is certainly an important research problem. Scalability is important for web scale problems. The Local Ranking method could be used for large scale web expert search: retrieve a moderate number (e.g., 20k) of top relevant pages from a traditional search engine and run CoDiffusion. The running time depends on the number of relevant pages, but not the size of the web collection in the index. In our current implementation, did not optimize the performance using multithreading, multicore, MapReduce or sampling techniques. There is room to further improve the running speed.

## VII. CONCLUSIONS

This paper proposes a new link analysis ranking approach for co-ranking authors and documents respectively in their social and citation networks. Starting from the PageRank paradigm as applied to both networks, presumably exploiting the mutually



| cs-id | title | authors | year | cite |
|---|---|---|---|---|
| 364205 | Learning Bayesian Networks: The Combination of Knowledge and Statistical Data | David Heckerman, Dan Geiger, David Chickering | 1994 | 351 |
| 142690 | Bounds on the Sample Complexity of Bayesian Learning Using Information Theory and the VC Dimension | David Haussler, Michael Kearns, Robert Schapire | 1992 | 85 |
| 124084 | Efficient Distribution-free Learning of Probabilistic Concepts | Michael J. Kearns, Robert E. Schapire | 1993 | 115 |
| 25286 | Bagging Predictors | Leo Breiman | 1996 | 657 |
| 384587 | Reinforcement Learning: Introduction | Richard Sutton | 1998 | 614 |
| 48796 | An Information-Maximization Approach to Blind Separation and Blind Deconvolution | Anthony J. Bell, Terrence J. Sejnowski | 1995 | 491 |
| 41366 | Stacked Generalization | David H. Wolpert | 1992 | 367 |
| 527057 | Optimization by Simulated Annealing | S. Kirkpatrick | 1993 | 1527 |
| 25887 | Mining Association Rules between Sets of Items in Large Databases | Rakesh Agrawal, Tomasz Imielinski, Arun Swami | 1993 | 921 |
| 20336 | Generalized Additive Models | Trevor Hastie, Robert Tibshirani | 1995 | 450 |
| 123646 | Experiments with a New Boosting Algorithm | Yoav Freund, Robert E. Schapire | 1996 | 500 |
| 528249 | Hierarchical Mixtures of Experts and the EM Algorithm | Michael I. Jordanand Robert A. Jacobs | 1993 | 472 |
| 543817 | The Strength of Weak Learnability | Robert E. Schapire | 1990 | 273 |
| 63435 | Systematic Nonlinear Planning | David McAllester and David Rosenblitt | 1991 | 226 |
| 434739 | Bayesian Interpolation | David J.C. MacKay | 1991 | 244 |

reinforcing relationship between authors and their co-authors. Experiments on a real world data set suggest that Co-Ranking is more satisfactory than counting the number publications or the total number of citations a given scientist has received. Also, it appears competitive with the PageRank algorithm as applied to the social network only.

(1) A larger empirical evaluation could be carried out to compare the Co-Ranking framework with other methods and find out, on which inputs it performs unsatisfactorily

(2)A formal analysis of the properties of the new Co-Ranking framework is required, including the effect of parameters m, n, k, λ on the ranking results, speed of convergence, stability, etc. It is also interesting to try to bring it into correspondence with the existing general frameworks for link based ranking. Expect there to be interesting interconnections with the HITS algorithm and its variations, if authors are viewed as authorities and documents are viewed as hubs.

## VIII. FUTURE WORK

(1) A larger empirical evaluation could be carried out to compare the Co-Ranking framework with other methods and find out, on which inputs it performs unsatisfactorily

(2) A formal analysis of the properties of the new Co-Ranking framework is required, including the effect of parameters m, n, k, _ on the ranking results, speed of convergence, stability, etc. It is also interesting to try to bring it into correspondence with the existing general frameworks for link based rankings (see e.g. [7]). Expect there to be interesting interconnections with the HITS algorithm and its variations, if authors are viewed as authorities and documents are viewed as hubs

(3) Other ways shall be explored for coupling random walks other than the one suggested in this paper. Several possibilities have been deemed unsatisfactory, however, presumably, the (m, n, k, ) - setting does not exhλust all meaningful ways to do that. Studying the effect of introducing different AD AD may serve as a starting point

(4) Presumably, the framework can be generalized for co-ranking entities of several types. Even for the case of two types, its applications are not limited to co-ranking authors and documents either.

## REFERENCES

[1]L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report Paper SIDL-WP-1999-0120 (version of 11/11/1999), Stanford Digital Library Technologies Project, 1999.

[2]Z. Gy¨ongyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In M. A. Nascimento, M. T. ¨ Ozsu, D. Kossmann, R. J. Miller, J. A. Blakeley, and K. B. Schiefer, editors, Proceedings of the Thirtieth International Conference on Very Large Data Bases (VLDB), pages 576–587, 2004.

[3]H. Yang, I. King, and M. R. Lyu. NHDC and PHDC: Non-propagating and propagating heat diffusion classifiers. In Proceedings of the 12th International Conference on Neural Information Processing (ICONIP), pages 394–399, 2005.

[4]Hongbo Deng1, Jiawei Han, Michael R. Lyu, and Irwin King Modeling and Exploiting Heterogeneous Bibliographic Networks for Expertise Ranking JCDL'12, June 10–14, 2012, Washington, DC, USA. ACM 978-1-4503-1154-0/12/06

[5]H. Yang, I. King, and M.R. Lyu, "Diffusionrank: A Possible Penicillin for Web Spamming," Proc. Ann. Int'l ACM SIGIR Conf Research and Development in Information Retrieval, pp. 431-438, 2007.

[6]D. Zhou, S. Orshanskiy, H. Zha, and C. Giles, "Co-Ranking Authors and Documents in a Heterogeneous Network," Proc. Int'l Conf. Data Mining (ICDM), pp. 739-744, 2007

[7]Ziyu Guan, Gengxin Miao, Russell McLoughlin, Xifeng Yan,and Deng Cai, "Co-Occurrence-Based Diffusion for Expert Search IEEECS Log Number TKDE-2011-06-0353.Digital Object Identifier no. 10.1109/TKDE.2012.49

[8]Nidhi Grover "Comparative Analysis Of Pagerank And HITS Algorithms" International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, October - 2012 ISSN: 2278-0181

[9]R.I. Kondor and J. Lafferty, "Diffusion Kernels on Graphs and Other Discrete Input Spaces," Proc. 19th Int'l Conf. Machine Learning (ICML), pp. 315-322, 2002

[10]X. Liu, W.B. Croft, and M. Koll, "Finding Experts in Community-Based Question-Answering Services," Proc. ACM Conf. Information and Knowledge Management (CIKM), pp. 315-316, 2005

[11]H. Ma, H. Yang, M.R. Lyu, and I. King, "Mining Social Networks Using Heat Diffusion Processes for Marketing Candidates Selection,"

Proc. ACM Conf. Information and Knowledge Management (CIKM), pp. 233-242, 2008

[12]C. Macdonald and I. Ounis, “Voting for Candidates: Adapting Data Fusion Techniques for an Expert Search Task,” Proc. ACM Conf. Information and Knowledge Management (CIKM), pp. 387-396, 2006.

[13]N. Craswell, A.P. de Vries, and I. Soboroff, “Overview of the Trec 2005 Enterprise Track,” Proc. Text Retrieval Conf. (TREC), 2005.

[14]K. Balog and M. De Rijke, “Associating People and Documents,” Proc. IR Research, 30th European Conf. Advances in Information Retrieval (ECIR), pp. 296-308, 2008