*International Journal of*
*Emerging Research in Management &Technology*
*ISSN: 2278-9359 (Volume-3, Issue-11)*

**Research Article** | **November 2014**

# Expert Web Search Engine Using Associated Keywords

**Narender Yevagani, J. Pavan Kumar**
Department of Computer Science and Engineering
Malla Reddy Engineering College (Autonomous), Telangana, India

***Abstract—***

***Search Engines are highly popular in the digital information era. Customized and Expert search is a regular practice of computer users on the web. The user friendly search is taking major role in various domains including Medical, Enterprise, Education, Social and Government etc . The performance of the Search always depends on the way the query is supplied to the system. Due lack of conciseness and vocabulary the scope of query grows into millions of matches. In this paper we studied and examined the way of an expert user behave on the web to find exactly what he required from the huge repositories. Mainly there are two challenges in this context the quality and correctness of the results varies due to dynamic web pages added from time to time, web pages are usually vague and ambiguous due to non specific queries. In this paper these problems are being addressed with associated terms through co-occurrence list. To assess relevance and reputation of a person name for a query topic the co-occurrence list is used. A hyper graph is used in this paper to maintain co-occurrence relationship, on which a heat diffusion based ranking algorithm is used. This algorithm is effective for retrieving experts and outperforms baseline algorithms significantly.***

***Keywords— Web Search, Co-occurrence list, Keyword search, Diffusion, Associated words, expert search***

## I. INTRODUCTION

In traditional organizational expert search, relevance is the major concern. However, considering the challenges mentioned above, we also need to consider a name's reputation for a query topic as well as the trustworthiness of data sources. We suspect the relevance and reputation can be captured by the large amount of keyword-name and name-name co-occurrences on the web. Using a large amount of co-occurrence information, noises could be suppressed since noisy co-occurrences would not appear frequently on the web. The problem in can thus be alleviated because Ana Ivanovic probably does not co-occur frequently with salient swimmers. In particular, we aim to address the new challenging issues by leveraging the linkage of experts exhibited on the web: 1) Relevance. Related experts should co-occur frequently on many web pages with the keywords in the query. 2) Reputation. Related experts should co-occur frequently with other people related to the query, regardless of whether they are experts or not. For example, a salient researcher could be co-mentioned with other researchers in his/her research areas many times; a senior user in an online forum would actively pursue threads for which he/she has expertise and co-occur with many other users. 3) Trustworthiness. Related experts tend to occur in high-quality web pages. This existing system has lot of disadvantages including 1. Huge amount of information is displayed from different number of people. 2. There is no perfect query answering environment. 3. Here there is no expert identification methodology. 4. Some pages are displayed irrelevant to current story. There are major challenges while applying these concepts 1) Compared to an organization's repository, ordinary web pages could be of varying quality and full of noises. Shows examples of noises from a news page of CNN, i.e., links to popular news stories and advertisements, which are usually irrelevant to the current story. 2) The expertise evidences scattered in web pages are usually vague and ambiguous. The second observation could be true for many domains, since humans are socialized and social activities shall be reflected on the web. Following these observations, we propose to model the co-occurrence relationships among people names and words in a heterogeneous hyper graph where web pages are treated as hyper edges with Page Rank scores as their weights. Then, we develop a novel heat diffusion model on the hyper graph. Based on this model, an expert ranking algorithm, called Co-occurrence Diffusion (Co-Diffusion for short), and is developed. Given a query, we treat keywords in the query as heat sources and perform heat diffusion. Names with the highest heat scores are returned. Intuitively, people who have strong connection with the query (i.e., frequently co-occur with query keywords and frequently co-occur with other people related to query keywords in high-quality pages) will be ranked high. Intrinsically, Co-Diffusion aggregates evidences collected from different web pages. This aggregation could be a good remedy for noises in web data. Co-Diffusion complements traditional language model-based methods, if it applies their relevance scores. The proposed system has the following advantages over existing search engines like : Diffusion model defines the experts. Experts provide the quality web pages information. Here we display strong connection query related results as output content.

## II. LITERATURE SURVEY

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it. Expert search gained increasing attention from both industry and academia. The TREC enterprise tracks [4] boomed research work on organizational expert search [2], [8] [12]. Variant expert search problems were also identified and addressed in other domains such as question answering [4], online forums [39] and academic

 

society [9], [7]. Expert search is a growing research area. Early approaches for expert search involve building a knowledge base which contains the descriptions of people's skills within an organization [10]. However, creating a knowledge base manually is time consuming and laborious. Therefore, automatic approaches have been developed for building people profiles [11], [6]. Expert search became a hot research area since the start of the TREC enterprise track [1] in 2005. A lot of studies were dedicated to organizational expert search. Balog et al. proposed a language model framework for expert search [7]. Their Model 1 is equivalent to a profile-centric approach where text from all the documents associated with a person is amassed to represent that person. Their Model 2 is a ocument-centric approach which first computes the relevance of documents to a query and then accumulates for each person the relevance cores of the documents that are associated with the person. This process was formulated in a generative probabilistic model. Balog et al. howed that Model 2 outperformed Model 1 [5] and it became one of the most prominent methods for expert search. In their following work, Balog et al. tried to apply and refine their language model on a smaller data set comprising multilingual data crawled from Tilburg University's website [4] Researchers have investigated using additional information to boost retrieval performance, such as PageRank, indegree, and URL length of documents [8], person-person similarity [9], internal document structures that indicate people's association with document content [6], query expansion and relevance feedback using people names [10], [7], nonlocal evidence [8], [11], proximity between occurrences of query words and people names [9], [12]. Besides language models, other methods have been proposed for organizational expert retrieval. Macdonald and Ounis proposed a method based on voting and data fusion techniques [12]. Serdyukov et al. modeled associations between people and documents as a bipartite graph and performed probabilistic random walks to find relevant experts [2]. Fang et al. proposed a relevance-based discriminative learning framework for expert search [8]. Many other methods for organizational expert search were proposed during TREC Enterprise tracks. There are other expert retrieval problems. Balog and de Rijke studied the problem of finding similar experts, given example experts [5]. Zhang et al. studied characteristics of online forums and tested using link analysis methods to identify users with high expertise [9]. Liu et al. studied expert finding in community-based question answering websites and treated it as an IR problem [4]. Mimno and McCallum used topic modeling to address the problem of matching papers with reviewers [12]. Later Karimzadehgan et al. addressed this review assignment problem based on matching of multiple aspects of expertise [2], [11]. Deng et al. explored using language modeling and a topic-based model for expert finding in the DBLP bibliography data [9]. Zhou et al. proposed coranking authors and their publications using coupled random walks [4]. Finally, our work is also related to heat diffusion on graphs. In real world, heat diffuses in a medium from position with a higher temperature to that with a lower temperature. The idea of heat diffusion was extended to the discrete graph setting, with applications such as dimension reduction [12], classification [3], antispamming [5], social network marketing [6] and online advertisement matching [1]. These studies considered diffusion in homogeneous graphs. In this paper, we develop a diffusion model based on heterogeneous hypergraphs for our expert search problem. Our approach to expert search assumes that we have a heterogeneous document repository, such as a corporate intranet, containing a mixture of different document types (e.g., technical reports, email discussion, web pages, etc). We assume that a document d in this collection is associated with a candidate ca, if there is a non-zero association a(d, ca) > 0. This association may capture various aspects of the relation between a document and a candidate expert; e.g., it may quantify the degree to which this document is representative of the candidate's expertise, or, vice-versa, it may capture the extent to which the candidate is responsible for the document's content. Forming document-candidate associations is a non-trival problem, which we consider in detail later in this paper . For now, we present our formal models assuming we have these associations.

## III. CO-OCCURRENCE BASED SEARCH

The second observation could be true for many domains, since humans are socialized and social activities shall be reflected on the web. Following these observations, we propose to model the co-occurrence relationships among people names and words in a heterogeneous hyper graph where web pages are treated as hyper edges with Page Rank scores as their weights. Then, we develop a novel heat diffusion model on the hyper graph. Based on this model, an expert ranking algorithm, called Co-occurrence Diffusion (Co-Diffusion for short), is developed. Given a query, we treat keywords in the query as heat sources and perform heat diffusion. Names with the highest heat scores are returned. Intuitively, people who have strong connection with the query (i.e., frequently co-occur with query keywords and frequently co-occur with other people related to query keywords in high-quality pages) will be ranked high. Intrinsically, Co-Diffusion aggregates evidences collected from different web pages. This aggregation could be a good remedy for noises in web data. Co-Diffusion complements traditional language model-based methods, if it applies their relevance scores.

### *A. Advantages*

*i.)* Diffusion model defines the experts.
*ii.)* Experts provide the quality web pages information.
*iii.)* Here we display strong connection query related results as a output content.

### *B. System Modules*

*i.)* Creation of forum and implement the search engine
*ii.)* Heat Diffusion on Heterogeneous Hyper graphs
*iii.)* Diffusion Model
*iv.)* Expert Search

 

*v.)* Global ranking versus Local ranking
*vi.)* Performance evolution results.

### *i. Creation of forum and implement the search engine*

An Internet forum, or message board, is an online discussion site where people can hold conversations in the form of posted messages. A discussion forum is hierarchical or tree-like in structure: a forum can contain a number of sub forums, each of which may have several topics. Within a forum's topic, each new discussion started is called a thread, and can be replied to by as many people as so wish. Renlifang1 is an object level search engine which allows users to query about people, locations, and organizations and explore their relationships. It employs entity extraction and relation extraction techniques. The major technique used in search engines like Renlifang is to extract structural information about entities and their relationships by deep-parsing web pages.

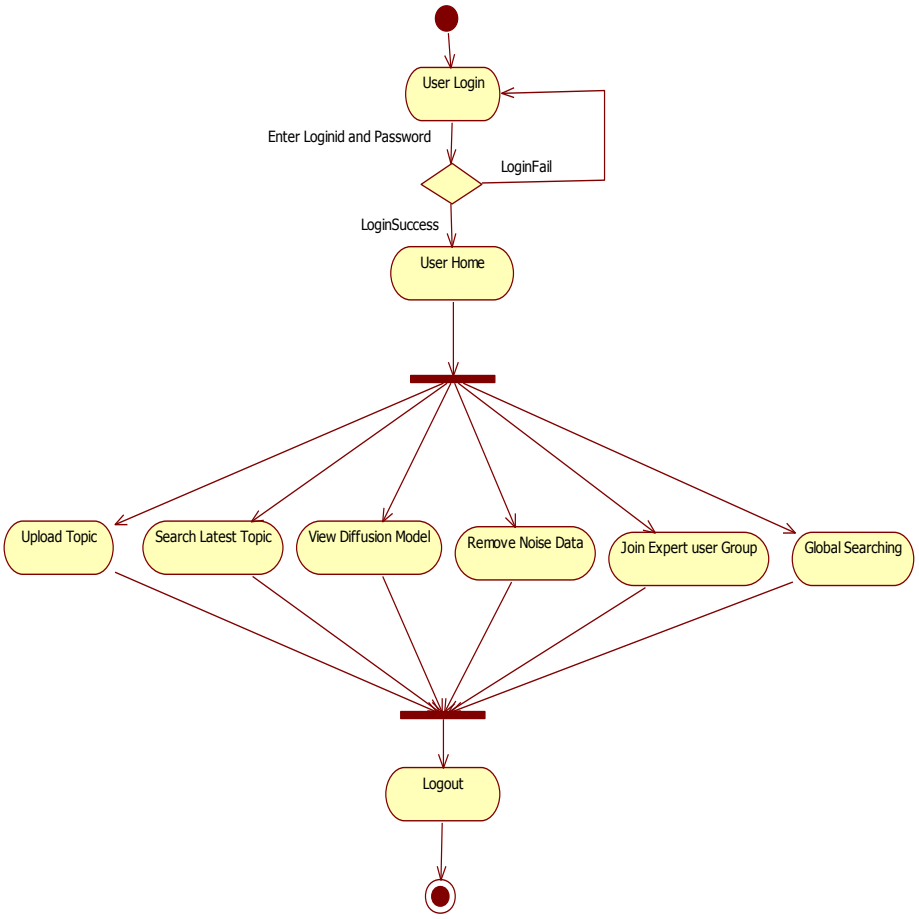


Fig 1. Co-occurrence based diffusion search by expert users Activity diagram

### *ii. Heat Diffusion on Heterogeneous Hyper graphs*

In a hypergraph, each edge i.e hyperedge can connect two or more vertices. $G=(V,E)$ be a hypergraph with vertex set $V$ and edge set $E$. A hyperedge $e \varepsilon E$ can be regarded as a subset of vertices. $e$ is said to be incident with a vertex $v$ if $v \varepsilon e$. Each hyperedge **e** is associated with a weight denoted by $w(e)$. In our problem setting, there are three types of objects: ***people*** (names), ***words***, and ***webpages***, denoted by $P$, $W$, and $D$, respectively. By the co-occurrence relationships among $P$ and $W$ established by ***webpages***, we can construct a heterogeneous hypergraph $G_{P,W}=(V,E)$ where $V$ contains vertices representing all the people and words and each $e \varepsilon E$ corresponds to a ***webpage***. A toy example is shown in Fig. 2. $W(e)$ is the PageRank score of e's corresponding webpage. The problem is, 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.

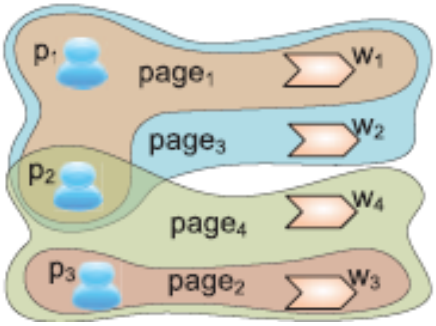


Fig 2.Heat Diffusion on Heterogeneous Hyper Pgraph

 

The aimed method using heat diffusion to address this ranking problem. Let **Vp** and **Vw** represent the vertex sets corresponding to people and words, respectively. Consequently, V = Vp U Vw. Let Hp be a |Vp|X|E| weighted incidence matrix where an entry ***Hp(v,e)= wtv;e*** if **vεe** (**vεVp**) and **0** otherwise. **Hw** is defined similarly for Vw. **wtv;e** reflects the connection strength between object v and page e. We set Hpðv; eÞ to the number of times person v appears in page e and set Hwðv; eÞ to the TF-IDF score of word v in e. The degree of a vertex v is defined as:

$$d(v) = \begin{cases} \sum_{e \in E} w(e) H_p(v,e) & v \in V_p \\ \sum_{e \in E} w(e) H_w(v,e) & v \in V_w. \end{cases} \quad (1)$$

The degree of a hyperedge is defined as

$$\delta(e) = \delta_p(e) + \delta_w(e) \quad (2)$$

Where $\delta_p(e) = \sum_{v \varepsilon Vp} H_p(v,e)$ and $\delta_w(e) = \sum_{v \varepsilon Vp} H_p(v,e)$ We define $f^p_i(t)$ and $f^w_i(t)$ to be the heat of vertex iεVp and that of vertex jεVw at time t, respectively. Let $f^p(t)$ and $f^w(t)$ be the heat distribution vectors at time t with sizes $|V^p|$X1and $|V^m|$X1, respectively. The initial heat distribution is represented by $f^p(0)$ and $f^w(0)$. Then, the problem is to derive the heat distribution at time t ($f^p(t)$ and $f^w(t)$ )given an initial distribution at time 0 ($f^p(0)$ and $f^w(0)$. In other words, we can set query objects (people and/or words) as heat sources and rank other objects according to the heat distribution at time t, which reflects the affinity between the objects and heat sources. This is a general ranking model. In our problem, words are queries and we need to get the ranking of people.

### *iii. Diffusion Model*

Heat diffuses in a medium from positions with higher temperatures to those with lower temperatures. The most important property of heat diffusion is that the heat flow rate at a point is proportional to the second order derivative of heat with respect to the space at that point. Different medium have different thermal conductivity coefficients. The diffusion model is constructed as follows: At time t, each vertex i € V will receive an amount of heat from its neighbours.

$$\frac{\partial f(\mathbf{x},t)}{\partial t} = \gamma \nabla^2 f(\mathbf{x},t) \quad (3)$$

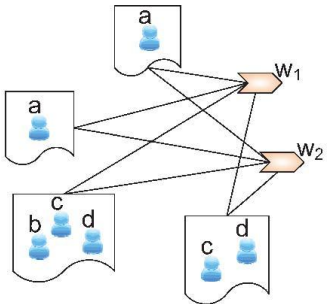


Fig 3. A toy problem which illustrates the effect of heat normalization term d'(v) for people

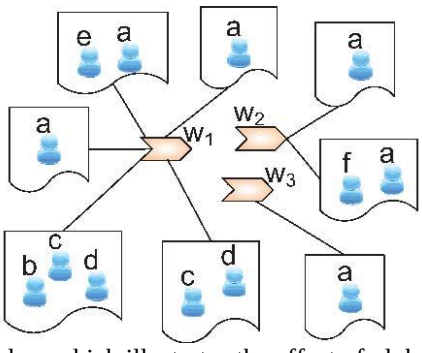


Fig 4.A toy problem which illustrates the effect of global normalization

### *iv. Expert Search*

**A**n expert should expose himself/herself more frequently than non experts. Therefore, we consider d(v) as a factor in d(v) for a name. Another characteristic of experts is that they tend to co-occur with many different people on the web, e.g., a professor would co-occur with many students and other professors; a senior forum user would actively answer questions for other users and consequently co-occurs with many different users.

### *v. Global ranking versus Local ranking*

There are two possible schemes to implement proposed algorithm: 1) we perform "Model Construction" on the entire web collection and for each query we only need to perform the "Diffusion and Ranking". In other words, the first phase of Algorithm needs to be done only once. Then, the constructed model is used for all queries. We call this scheme Global Ranking; 2) we first obtain related web pages for a query by querying the web collection.

 

### *vi. Creation of forum and implement the search engine*

An Internet forum, or message board, is an online discussion site where people can hold conversations in the form of posted messages. A discussion forum is hierarchical or tree-like in structure: a forum can contain a number of sub forums, each of which may have several topics. Within a forum's topic, each new discussion started is called a thread, and can be replied to by as many people as so wish. Renlifang1 is an object level search engine which allows users to query about people, locations, and organizations and explore their relationships. It employs entity extraction and relation extraction techniques. The major technique used in search engines like Renlifang is to extract structural information about entities and their relationships by deep-parsing web pages.

## IV. PERFORMANCE EVOLUTION RESULTS

The diffusion process employed in Algorithm only sets query keywords as heat sources (i.e., queries). This can overly emphasize word-name diffusion and reduce the effect of name-name diffusion. Here, we propose two reranking algorithms to refine the ranking results by setting top ranked people names as heat sources (i.e., queries), in order to boost reputable names for the query. The first reranking algorithm is named One-Time Re- Ranking. The idea is that we set top k names from the ranking result generated by CoDiffusion as queries and invoke CoDiffusion (without global normalization) a second time. The intuition is that the top k names can be regarded as expert candidates and we could boost reputable experts by diffusing heat from these candidates. In the second reranking algorithm, we use an iterative process to gradually refine ranking results: initially we choose top k names from the result of CoDiffusion and use pages which contain at least two names in the top k names to build the diffusion model. Then, we set these k names as queries and invoke CoDiffusion (without global normalization); in the jth iteration we perform the same process with top k-(j-1)k0 names from the last iteration, where k0 is a small value (e.g., 50). By the second algorithm, we try to perform more and more focused diffusion in the community to find reputable experts. The second algorithm is named Iterative Re-Ranking. We summarize the two algorithms, respectively. For Iterative Reranking, we discard names other than names in Top to better focus on top ranked names. We use the corresponding ranking scores outputted by CoDiffusion as query weights. In this way, the final ranking result will not deviate too much from the original one.

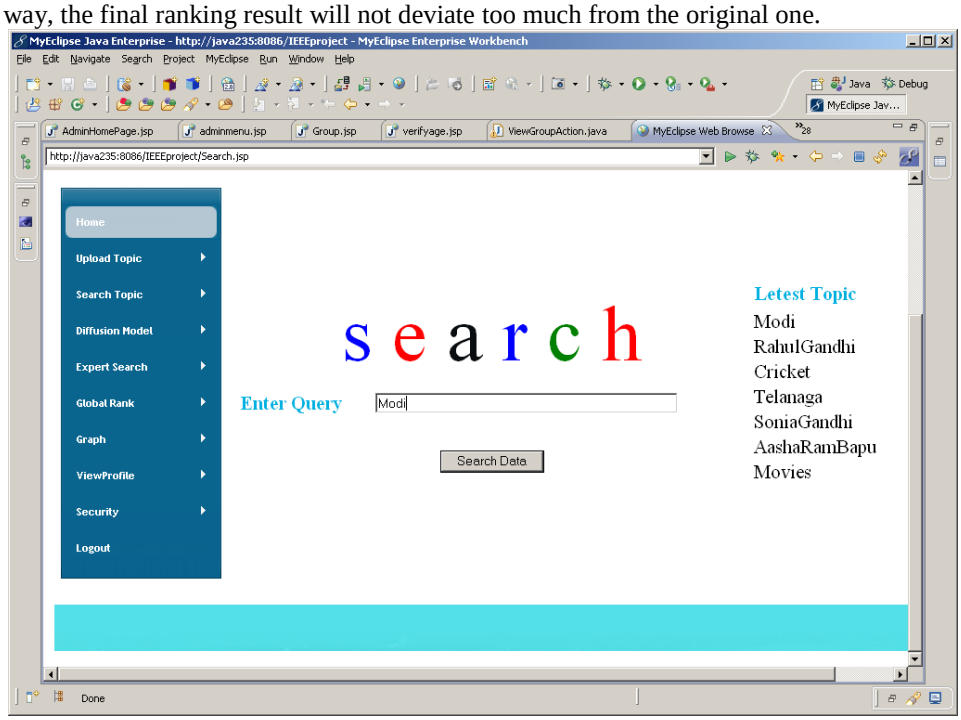


Fig 5. Expert Search Interface

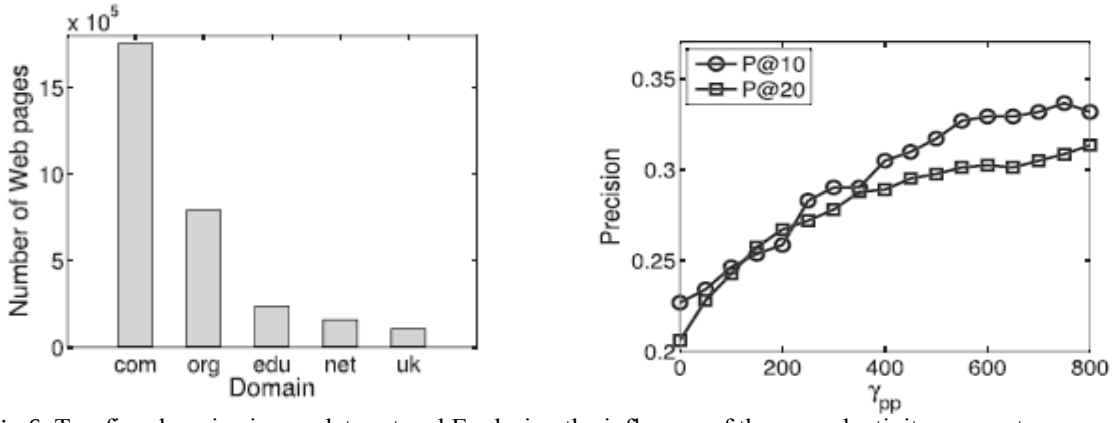


Fig 6. Top five domains in our data set and Exploring the influence of three conductivity parameters $\gamma_{pp}$

 

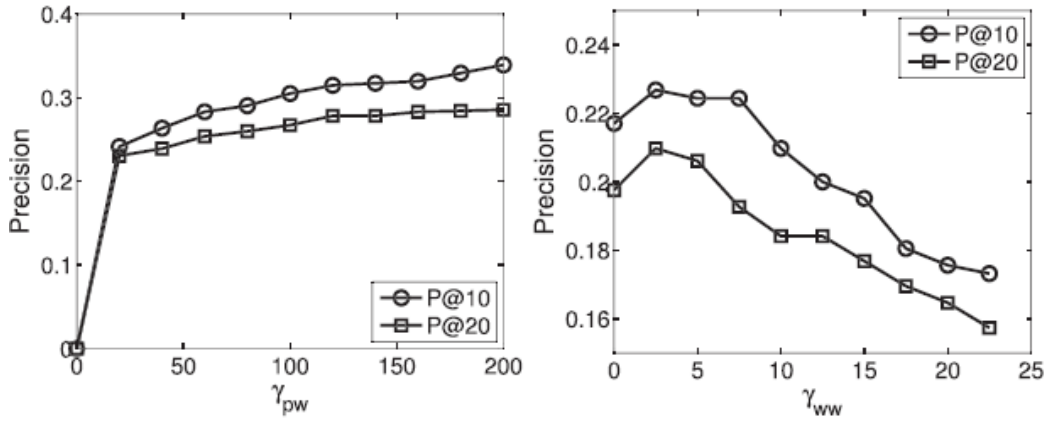


Fig 7. $\gamma_{pw}$ , and $\gamma_{ww}$ on the performance of CoDiffusion. For each parameter,the other two parameters are fixed at 1. Results are averaged over all the 50 queries

## V. CONCLUSION

In this paper, we studied a general expert search problem on the web. We proposed not to deep-parse webpages for expert search. Instead, it is possible to leverage cooccurrence relationships such as name-keyword co-occurrences and name-name co-occurrences to rank experts. A ranking algorithm called CoDiffusion was developed based on this concept. CoDiffusion adopts a heat diffusion model
on heterogeneous hypergraphs to capture expertise information encoded in these co-occurrence relationships. Experiments on ClueWeb09 and two benchmark data sets consisting of research queries demonstrated that CoDiffusion outperformed the baseline algorithms significantly. Experiments on conductivity coefficients verified that cooccurrences were indeed useful. We also explored queries other than research related topics and showed that CoDiffusion could return good results and outperform baselines.

## REFERENCES

[1] Artiles, J. Gonzalo, and S. Sekine, "Weps 2 Evaluation Campaign: Overview of the Web People Search Clustering Task," Proc. Second Web People Search Evaluation Workshop (WePS '09),2009.

[2] K. Balog, L. Azzopardi, and M. de Rijke, "A Language Modeling Framework for Expert Finding," Information Processing & Management,vol. 45, no. 1, pp. 1-19, 2009.

[3] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting Spam Web Pages through Content Analysis," Proc. Int'l Conf. World Wide Web (WWW), 2006.

[4] S. P. Serdyukov and D. Hiemstra, "Being Omnipresent to be Almighty: The Importance of the Global Web Evidence for Organizational Expert Finding," Proc. SIGIR Workshop Future Challenges in Expertise Retrieval (fCHER), pp. 17-24, 2008.

[5] J. Tang, A. Fong, B. Wang, and J. Zhang, "A Unified Probabilistic Framework for Name Disambiguation in Digital Library," IEEE Trans. Knowledge and Data Eng., vol. 24, no. 6, pp. 975-987, June 2012.

[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] Y. Fu, W. Yu, Y. Li, Y. Liu, M. Zhang, and S. Ma, "THUIR at Trec 2005: Enterprise Track," Proc. Text Retrieval Conf. (TREC), 2005.

[8] M. Karimzadehgan and C. Zhai, "Constrained Multi-Aspect Expertise Matching for Committee Review Assignment," Proc. ACM Conf. Information and Knowledge Management (CIKM), pp. 1697-1700, 2009.

[9] H. Deng, I. King, and M.R. Lyu, "Formal Models for Expert Finding on DBLP Bibliography Data," Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 163-172, 2009.

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

[11] M. Yoshida, M. Ikeda, S. Ono, I. Sato, and H. Nakagawa, "Person Name Disambiguation by Bootstrapping," Proc. Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 10-17, 2010.

[12] D. Yimam-Seid and A. Kobsa, "Expert-Finding Systems for Organizations: Problem and Domain Analysis and the Demoir Approach," J. Organizational Computing and Electronic Commerce, vol. 13, no. 1, pp. 1-24, 2003.

 