International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

An Optimized Ranking Strategy for Expert Search on the Web with NLP Techniques

Kiran G. Shinde[1], S. B. Natikar[2]

[1, 2]Computer Engineering Department, VACOE, Pune University, Maharashtra, India

Abstract: ***The most important thought of this strategy is to find experts for variety of domains on the web, where large numbers of WebPages and people names are considered. It has mainly two difficult issues: WebPages could be of unreliable quality and full of noises; the expertise evidences scattered in WebPages are usually formless and uncertain. We propose to control the large amount of co-occurrence information to evaluate relevance and reputation of a person name for a query topic. The objective is to design a system providing functionality of the expert search engine.NLP techniques can be applied for the same with name queries. Performance is optimized by effective crawling and deep parsing of web pages in order to adjust the association scores among people names and words. Global occurrences of experts are exercised so as to support the accuracy and relevancy of results. The proposed system also tries to boost performance by user rating based on user feedback. Finally we propose a unified approach for Person name Extraction where crawled data from web is applied to module which uses 3 class models which is used for building code for developing sequence models.***

Keywords: Co-occurrence, Expert Search, NLP Techniques, Sequence Models

1. Introduction

Expert search gained increasing attention from both industry and academia. The TREC enterprise tracks boomed research work on organizational expert search. Variant expert search problems were also identified and addressed in other domains such as question answering, online forums and academic society. Various expert search problems were also acknowledged and applied in other fields such as question answering [4], online environments [5] and educational society [6], [7], [8]. A feature shared by many of the models proposed for ranking people with respect to their expertise on a given topic is their reliance on associations between people and documents. Ex..If someone is strongly associated with an important document on a given topic, this person is more likely to be an expert on the topic than Someone who is not associated with any documents on the topic.

A. Purpose:

To develop a system that gives optimal solution for current variety of daily life issues. The proposed heat distribution based ranking algorithm uses co-occurrence configuration that is modeled using a hypergraph. Query keywords are observed as heat sources, and individual name which has well-built relation with the query (i.e., regularly co-occur with query keywords and co-occur with further names associated to query keywords) will get the majority of the heat, so as to rank high. To optimize the performance of existing system using multithreading, multicore and map reduce or sampling techniques

B. Objective of the System

We examine a general expert search problem: finding experts on the web, where large numbers of WebPages and people names are considered. It has mainly two difficult issues: WebPages could be of unreliable quality and full of noises; the expertise evidences scattered in WebPages are usually formless and uncertain. We propose to control the large amount of co-occurrence information to evaluate relevance and reputation of a person name for a query topic. The objective is to design a system providing functionality of the expert search engine.NLP techniques can be applied for the same with name queries. The system should operate in the multithreading environment as well. We also try to boost performance by reranking based on name pseudo relevance Feedback.

2. Related Work

Users of the internet often have the urge to discover biographies and data of people of interest. For celebrity biographies and facts, Wikipedia is the first choice for number of users. However, Wikipedia can only give information for personalities for the reason that it has its neutral point of view (NPOV) editorial policy. Expert search is a emerging research area. Prior approaches for expert search engross constructing a knowledge base that includes the descriptions of candidate's abilities within an organization [9].A lot of studies were devoted to organizational expert search.

Aardvarks facilitate users to ask a question, by direct message or email, text message or voice. Aardvark then forwards the question to the individual in the user's total network possibly capable of answering that question. In comparison with a conventional web search engine, where the difficulty lies in finding the accurate document to satisfy a user's information requirement, the confront in a social search engine like Aardvark lies in discovering the exact person to complete a user's information need.

Balog et al. put forwarded a language model framework for expert search [10]. Their Model 1 is similar to a profile-centric approach where text from all the documents associated with a person is amassed to represent that person. Their Model 2 provides a document-centric strategy which first computes the relevance of documents to a query and then accumulates for each person the relevance scores of the documents that are associated with the person. Generative probabilistic model formulated this process. Balog et al. showed that Model 2 performed better than Model 1 [10] and it turned out to be one of the most promising methods for

expert search. In their subsequent work, Balog et al. attempted to relate and refine their language model on a smaller data set containing multilingual data which is crawled from Tilburg University's website [10].

Expert finding, is a multidisciplinary problem that cross-cuts knowledge management, organizational analysis, and information retrieval. Recently, a number of expert finders have emerged; however, many tools are limited in that they are extensions of traditional information retrieval systems and exploit artifact information primarily.

The *Expert Locator*, developed within a live organizational environment, is a model-based prototype that exploits organizational work context. The system associates expertise ratings with expert's signaling behavior and is extensible so that signaling behavior from multiple activity space contexts can be fused into aggregate retrieval scores. Post-retrieval analysis supports evidence review and personal network browsing, aiding users in both *detection* and *selection*. During operational evaluation, the prototype generated high-precision searches across a range of topics, and was sensitive to organizational role; ranking *true* experts (i.e., *authorities*) higher than *brokers*. Researchers have examined using supplementary information to improve retrieval concert, such as Indegree, PageRank, and URL extent of documents [11], person-person similarity [2], internal document structures that indicate people's association with document content [12], query expansion and relevance feedback using people names[13], [14], nonlocal evidence [15], [16], proximity between occurrences of query words and people names [17], [18].In the context of organizational expert retrieval ,apart from language models, other methods have been proposed. Macdonald and Ounis projected a method based on voting and data fusion techniques [19]. Serdyukov et al. modeled associations between people and documents as a bipartite graph [20].Fang et al. proposed a relevance-based discriminative learning framework for expert search [21].Many other methods for organizational expert search were proposed during TREC Enterprise tracks.

Two benchmark data sets, W3C [22] and CSIRO [23], are the focus of the above organizational expert search works, which are formed after information extracted from the websites of World Wide Web Consortium and Commonwealth Scientific and Industrial Research Organization. Conversely, searching experts on the web is dissimilar from organizational expert search in that we recognize ordinary WebPages and people names.

There are other expert retrieval problems. Balog and deRijke studied the problem of finding similar experts, given example experts [24]. Zhang et al. worked on characteristics of online forums and tested using link analysis methods to recognize users with high expertise [5]. Liu et al. studied expert finding in community-based question answering websites and treated it as an IR problem [4].

Finally, our work is also related to heat diffusion on graphs. The concept of heat diffusion was modified to the discrete graph setting, with applications like dimension reduction [6], classification [26], social network marketing [27] and online advertisement matching [3]. These studies considered diffusion in homogeneous graphs. In this paper, we develop a diffusion model based on heterogeneous hypergraphs for our expert search problem.

3. Implementation

The technologically driven world in which we live in has increased the necessity for human interaction with system, particularly with computer-based system that are used to accomplish a vast variety of tasks with the aim of helping the user in achieving goal.

A. Heat Distribution on Hypergraphs

In a hypergraph, each edge (called hyperedge) can connect two or more number of vertices. Formally, let G = (V,E) be a hypergraph having vertex set V and edge set E. In our problem background, there are three kinds of objects: people (names), words, and WebPages, denoted by P, W, and D, respectively. By the co-occurrence association among P and W established by WebPages, we can construct a heterogeneous hypergraph. A toy example is shown in Fig. 1. W (e) is the Page Rank score of e's corresponding webpage. The problem is, given P, W, Gpw and query keywords from W, to rank P according to their expertise in the topic represented by the query

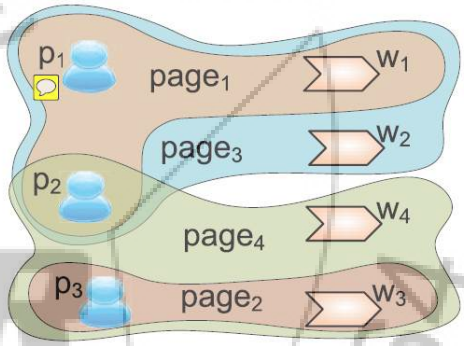

Figure 1: An example of heterogeneous hypergraph

In our problem, words are queries and we need to get the ranking of people.

B. Distribution Model

The perception behind the distribution model is as follows: by constructing the matrix L, we basically combined the co-occurrence information among people and words to imitate the correlation strength between each couple of objects. This aggregation could be supportive for handling the problem of noises on the web. After the creation of L, we disseminate heat from query keywords (i.e., (17)) on this aggregated structure. Intuitively, names having strong connection not only with query keywords but also with other related names and words will be ranked high.

C. Person Name Identification and Extraction

In order to form document-candidate associations, we need to be able to recognize candidates' occurrences within documents. In the TREC setting, a list of possible candidates is given, where each person is described with a unique person

id, one or more names, and one or more e-mail addresses. While this is a specific way of identifying a person, and different choices are also possible (e.g., involving social security number instead of, or in addition to, the representations just listed), nothing in our modeling depends on this particular choice.

We propose a unified approach for Person name Extraction where crawled data from web is applied to module which uses Stanford NER which is CRF Classifier which is used for building code for developing sequence models. Models build with Stanford NERare 4 class models, 7 class models and 3 class models

3 class	Location, Person, Organization
4 class	Location, Person, Organization, Misc
7 class	Time, Location, Organization, Person, Money, Percent, Date.

D. Multithreaded Environment and NLP

Above mentioned strategy delivers proper functionality but the issue remains for handling large amount of data. Again the Problem of scalability can be removed by providing mentioned algorithm with multithreading environment. When different threads in algorithm are independent of each other there multithreading can be applied to improve the running speed. The ranking algorithm can be optimized to deliver accurate functionality with improved speed. While considering the fact of improving association scores between documents and people the trustworthiness of resources is taken into account. The quality of WebPages we are relying on can be checked and page weight can be calculated with the help of improved NLP Techniques

E. Algorithm

I. Data preparation and Model Construction:

a) As we are going to apply the algorithm on information retrieved from the web pages so we first need-
- A huge collection of web pages D1, D2, D3…Dn.
- Domain list L1, L2, L3…Ln.
- Frequently occurred domain words d1,d2,d3…dn. So as to decide domain of that webpage.

b) Apply stemming and conflation algorithm on webpage contents to remove stop words and noisy data.
c) Pass these refined contents to named entity recognizer model which extracts list of persons, location and organization in that web page.
d) Decide domain of web page by matching domain words to query keywords.

II. Algorithm and Ranking:

A. We aggregate the co-occurrence information among people and words to reflect connection strength between each pair of objects.

1. We construct a hypaergraph Gp,w = (V,E)
   V is set of vertices representing all people and words
   Each e € E corresponds to a web page.
   Let Vp and Vw represent the vertex sets corresponding to people and words, respectively. ,
2. V = Vp Ú Vw.
   Let Hp be a [Vp] *[E] weighted incidence matrix where an entry Hp (v, e) = wt (v, e).
3. Hw is defined similarly for Vw. Wt (v, e) reflects the connection strength between object v and page e.
4. We set Hp(v,e) to the number of times person v appears in page e and
   Set Hw (v, e) to the TF-IDF score of word v in e. The degree of a vertex v is defined as

$$d(v) = \sum_{e \in E} w(e)Hp(v,e) \; v \in Vp$$

$$d(v) = \sum_{e \in E} w(e)Hw(v,e) \; v \in Vw$$

5. We construct matrix L such as

$$L = \begin{matrix} Lpp & Lpw \\ Lwp & Lww \end{matrix}$$

F. Time Complexity

Complexity of system is mainly based on web crawler and Information Extraction module. If web crawler module doesn't have maximum URLs to search variable then problem is NP- Hard, because it cannot be solved in polynomial time. If system is set to search "n" numbers of URL in web crawler module, then it will create "n" crawled pages. Information Extraction module will extract information from "n" crawled pages. So time complexity of system is Θ (n2).

4. Proposed Framework

Fig. 1 shows the framework of our approach. The proposed system framework is an enhancement to techniques introduced in [1].The main motive of proposed system is to identify experts and return search results within few time. New framework makes use of co-occurrence information and Hierarchical clustering.

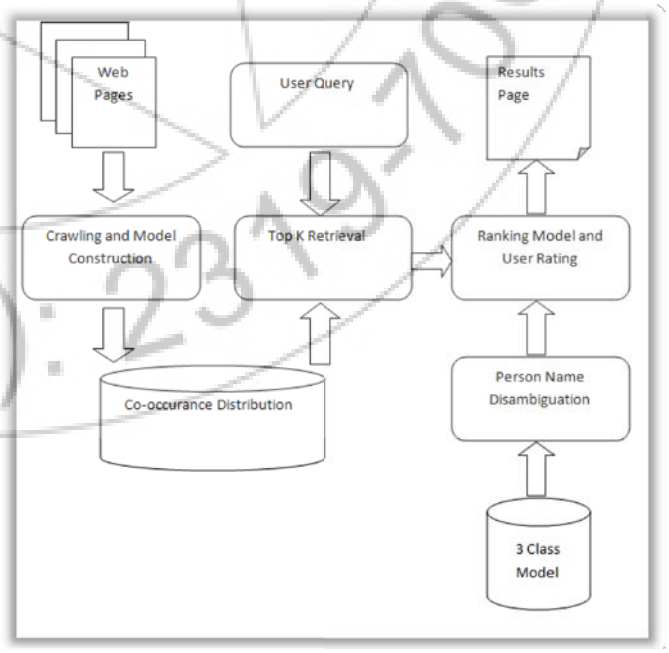

Figure 2: Proposed System Model (Framework)

5. Mathematical Model

1) Let S be a system that describes Expert Search Engine S={ ..... }.
2) Identify input as I
S= {I.....
I= {Q | where Q is input query string}
3) Identify output as O.
S= {I, O .....
O= { Ri | Ri is output strings that gives experts list of users interest.}
4) Identify the processes as P.
S= {I, O, P....
P= {WC, IE, AD}
* WC is Web Crawler
* IE is Information Extraction.
* AD is Association Distribution
5) WC= {URL, MAX,CP}
URL is uniform resource locator given as input to Web Crawler. MAX is maximum no of URL to be crawled
CP is output of Web Crawler which is Crawled web pages.
6) IE= {CP, NLP Techniques, Info}
CP is input which is Crawled web pages given to IE NLP rules are set of criteria to extract information Info is output of Information Extraction.
7) OP= {Info, AD, RR}
IE is extracted information given as input to Association Distribution
AD is ranking Algorithm given as input to extracted data
RR is Reranking on the data i.e. output of ranking Algorithm
8) Identify failure cases as F.
S = {I, O, P, F...
Failure occurs when O! =Ri
9) Identify Success cases as s
S= { I O, P, F, s...}
Success is defined as O= Ri
10) Identify initial condition IC
S= {I, O, P, F, s, IC}
There is no initial condition IC= {null}

6. Results and Discussion

A. Data Set

In order to perform and implement our work a rough data set is collected from World Wide Web that consists of ordinary web pages and people names. Our experimental datasets were extracted from the WebPages which is the result of web crawling module. The process for generating our experimental data sets is as follows: first we did a sequential scan through large number of WebPages to extract all the occurrences of author names, are used to find name occurrences. We discarded names which did not appear in those English pages. After this step we got distinct people names and pages, each of which contains at least one person name. We extracted and processed those pages using Stanford Core NLP provides a set of natural language analysis tools that takes raw English language text input and give the base forms of words like name of companies, people etc. and built index for them.

B Results and Performance measurement

We attempt to utilize the baseline algorithms [5] for performance measurement.. The first one, which is named Codiffusion, computes the total number of times a name emerges in pages that include all the query keywords. Occurrence in each page is weighted by the equivalent PageRank score. OMRN is the Optimized Multithreaded Ranking Algorithm that we exercised on two data sets Expertdb and Expertdb1.

Basketball	Algorithm	Java
Perry Jones	David Hilbert	James Gosling
Katie Karpowicz	Jaques Herband	Edward Showden
Kara Carmichael	J. Barkley Rosser	Pritesh Taral
Julian Zeng	S.C. Kleene	Steve Jobs
Jessi Langsen	David Rumelhert	Anupam Anand

Table 1.Top 5 names Returned by Algorithm for 3 queries

Three metrics are used for performance evaluation[5]: Precision@n (P@n), Mean Average Precision (MAP), P@n is the precision at rank n, which is defined as

Table 1: Performance Comparison of Expert Search algorithms

Algorithm	P@10	P@20	MAP
Expertdb			
Codiffusion	0.4125	0.3625	0.4977
OMRN	0.4925	0.4215	0.5255
Expertdb1			
Codiffusion	0.1542	0.1458	0.2647
OMRN	0.2552	0.2833	0.3323

$$P|@n = \frac{\text{No. Of Relevant Experts in top n results}}{n}$$

Average Precision is the average of precision scores after each correctly identified relevant expert:

$$AP = \frac{\sum P@i * Corr(i)}{\text{no. of correctly identified relevant experts}}$$

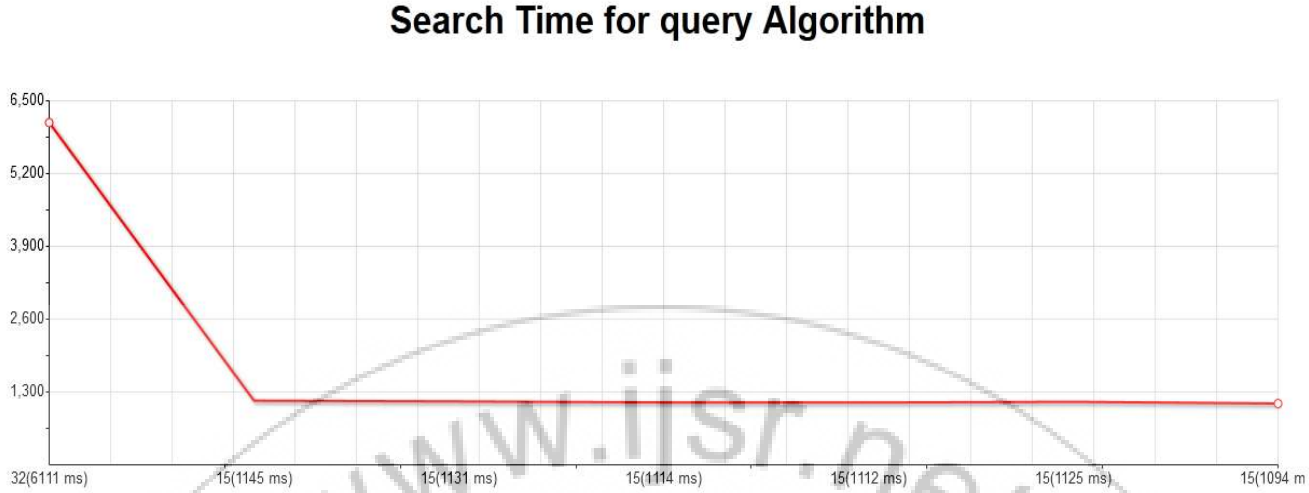

Figure 2: Graph showing Retrieval time for user query

A graph is plotted as number of results versus retrieval time (milliseconds) .The graph shows retrieval time taken for query "Algorithm".

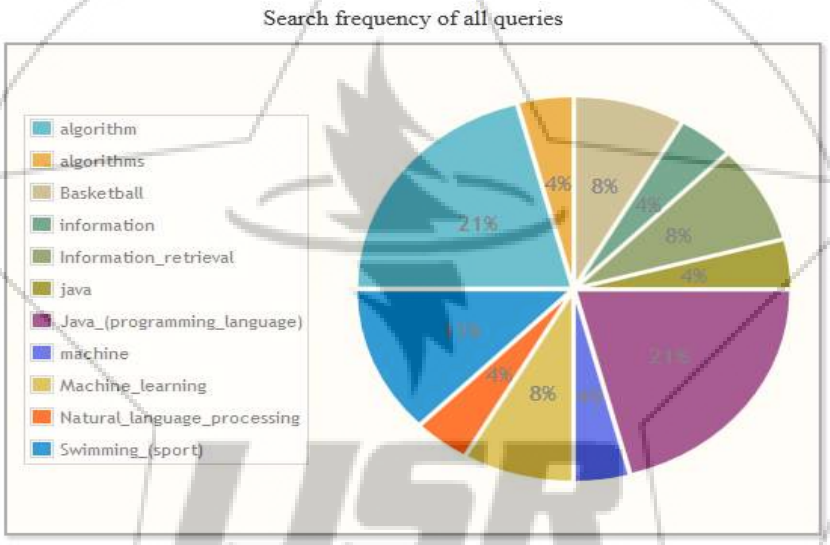

Figure 3: Search Frequency Of All Queries

7. Acknowledgment

I wish to express my sincere thanks and deep gratitude towards Dr. M. G Jadhav [Principal JSCOE] and my guide Prof. S. B. Natikar. For his guidance, valuable suggestions and constant encouragement in all phases. I am highly indebted to his help in solving my difficulties which came across whole Paper work. Finally I extend my sincere thanks to respected Head of the department Prof. S. G. Joshi and Prof. M C. Kshirsagar [P.G Co-Ordinator] and all the staff members for their kind support and encouragement for this paper. Last but not the least, I wish to thank my friends.

References

[1] J. Zhang, M.S. Ackerman, and L. Adamic, "Expertise Networks in Online Communities: Structure and Algorithms," Proc. Int'l Conf. World Wide Web (WWW), pp. 221-230, 2007..

[2] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, and A. van den Bosch, "Broad Expertise Retrieval in Sparse Data Environments," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 551-558, 2007

[3] H. Bao and E.Y. Chang, "Adheat: An Influence-Based Diffusion Model for Propagating Hints to Match Ads," Proc. Int'l Conf. World Wide Web (WWW), pp. 71-80, 2010.

[4] 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.

[5] Ziyu Guan, Gengxin Miao, Russell McLoughlin, Xifeng Yan, Member, IEEE, and Deng Cai, Member, IEEE) "Co-Occurrence-Based Diffusion for Expert Search on the Web" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 5, MAY 2013

[6] D. Mimno and A. McCallum, "Expertise Modeling for Matching Papers with Reviewers," Proc. 13th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 500-509, 2007..

[7] 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.
[8] 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
[9] P.R. Carlile, "Working Knowledge: How Organizations Manage What They Know," Human Resource Planning, vol. 21, no. 4, pp. 58- 60, 1998..
[10] K. Balog, L. Azzopardi, and M. de Rijke, "Formal Models for Expert Finding in Enterprise Corpora," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 43-50, 2006.
[11] J. Zhu, X. Huang, D. Song, and S. Ru¨ ger, "Integrating Multiple Document Features in Language Models for Expert Finding," Knowledge and Information Systems, vol. 23, no. 1, pp. 29-54, 2010.
[12] 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.
[13] C. Macdonald and I. Ounis, "Expertise Drift and Query Expansion in Expert Search," Proc. ACM Conf. Information and Knowledge Management (CIKM), pp. 341-350, 2007
[14] K. Balog and M. de Rijke, "Combining Candidate and Document Models for Expert Search," Proc. 17th Text Retrieval Conf. (TREC), 2008.
[15] K. Balog and M. de Rijke, "Non-Local Evidence for Expert Finding," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM), pp. 489-498, 2008..
[16] 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.
[17] 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..
[18] 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.
[19] 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
[20] P. Serdyukov, H. Rode, and D. Hiemstra, "Modeling Multi-Step Relevance Propagation for Expert Finding," Proc. ACM Conf. Information and Knowledge Management (CIKM), pp. 1133-1142, 2008.
[21] Y. Fang, L. Si, and A.P. Mathur, "Discriminative Models of Integrating Document Evidence and Document-Candidate Associations for Expert Search," Proc. 33rd Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 683-690, 2010.
[22] N. Craswell, A.P. de Vries, and I. Soboroff, "Overview of the Trec 2005 Enterprise Track," Proc. Text Retrieval Conf. (TREC), 2005.
[23] K. Balog, P. Thomas, N. Craswell, I. Soboroff, P. Bailey, and A.P. de Vries, "Overview of the Trec 2008 Enterprise Track," Proc. Text Retrieval Conf. (TREC), 2008.
[24] K. Balog and M. de Rijke, "Finding Similar Experts," Proc. Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 821-822, 2007.
[25] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," Neural Computation, vol. 15, no. 6, pp. 1373-1396, 2003.
[26] 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.
[27] 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

Author Profile

Kiran G. Shinde received the B.E. degree in Information Technology from Pune University; India in 2011.He is pursuing M.E.in Computer Engg. at University of Pune. His research interests include data mining, Information Retrieval, Web mining.

Siddaling B. Natikar received the B.E. and M.E. degrees in Computer Engineering from Vishweswaraya Technological University Belgaum, Karnataka, India. He is currently As Asst. Prof. at VACOE, Pune University, India. His research interests include Cloud Computing, Image processing, Information retrieval.