International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Expert Search Engine Using Co-Diffusion

Sayali Nikam[1], Trupti Raikar[2], Priyanka Tile[3], Prof.S.N.Bhadane[4]
[1,2,3]Student, [4]Guide Department of Computer Engineering, Pune Vidhyarthi Griha's College of Engineering, Nashik,
[1,2,3,4]University Of Pune , Maharashtra, India

Abstract— Expert search has been studied in different contexts, e.g., enterprises, academic communities. We examine a general expert search problem: searching experts on the web, where millions of web pages and thousands of names are considered. It has mainly two challenging issues: 1) web pages could be of varying quality and full of noises; 2) The expertise evidences scattered in web pages are usually vague and ambiguous. We propose to leverage the large amount of co-occurrence information to assess relevance and reputation of a person name for a query topic. The co-occurrence structure is modeled using a hyper graph, on which a heat diffusion based ranking algorithm is proposed. Query keywords are regarded as heat sources, and a person name which has strong connection with the query (i.e., frequently co-occur with query keywords and co-occur with other names related to query keywords) will receive most of the heat, thus being ranked high. Experiments on the ClueWeb09 web collection show that our algorithm is effective for retrieving experts and outperforms baseline algorithms significantly this work would be regarded as one step toward addressing the more general entity search problem without sophisticated NLP techniques.

Keywords— Expert search, Web mining, Hyper graph, Diffusion, Co-occurrence, Re-ranking.

I. INTRODUCTION

Expert search gained increasing attention from both industry and academia. Variant expert search problems were also identified and addressed in other domains such as question answering, online forums and academic society. Recently, the desire to find experts on a variety of daily life topics is increasing. We are observing a rising search paradigm that allows users to search for people who can answer their natural language questions. However, this system requires users to register and join a community. In contrast, the Web contains a huge amount of information about people (e.g. personal home pages, blogs, Web news). It is possible to build a powerful expert search engine by exploiting the information about people on the Web. In this paper we propose a general expert search problem: expert search on the Web, which considers ordinary Web pages and people names. This problem is different from organizational expert search and is more like Google where our goal is to return a list of experts with reasonable quality.

It has new challenges:

A. Compared to an organization's repository, ordinary Web pages could be of varying quality and full of noises.

B. The expertise evidences scattered in webpages are usually vague and ambiguous. In particular, we aim to address the new challenging issues by leveraging the linkage of experts exhibited on the web: 1) Relevance 2) Reputation 3) Trust worthiness

II. PROPOSED SYSTEM

A. System Architecture

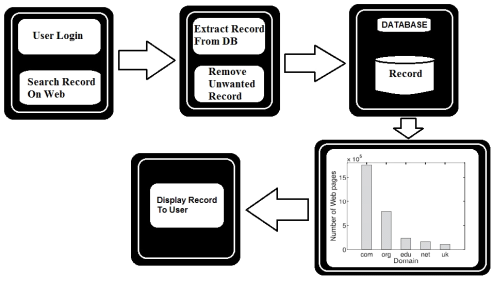

Fig. 1: System Architecture

1) Working Modules of Proposed System

a) Heterogeneous Hyper-graph: In a hyper graph, each edge (called hyper edge) can connect two or more vertices. In the

problem setting, there are three types of objects: people (names), words, and web pages, denoted by P, W, and D, respectively. By the co-occurrence relationships among P and W established by web pages, It can construct a heterogeneous hyper graph GP;W (V, E) where V contains vertices representing all the people and words and each e E corresponds to a webpage. A toy example is shown in w(e) is the Page Rank score of es corresponding webpage. The problem is, given P, W, GP,W and query keywords from W, to rank P according to their expertise in the topic represented by the query.

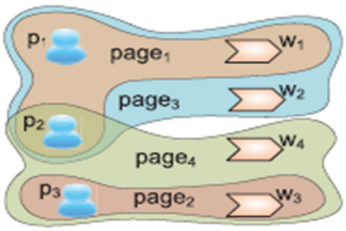

Fig. 2: An Example of Heterogeneous Hyper-graph

b) Diffusion: 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 neighbors.

c) Ranking: There are two possible schemes to implement our algorithm:

i. To perform Model Construction on the entire web collection and for each query, only need to perform the Diffusion and Ranking part in Algorithm 1. In other words, the first phase of Algorithm 1 needs to be done only once. Then, the constructed model is used for all queries. This scheme call as Global Ranking;

ii. First obtain related web pages for a query by querying the web collection.

- *Global Ranking*

  In Global Ranking the algorithm can diffuse heat to partially relevant or even irrelevant pages, while Local Ranking can perform more focused diffusion.

- *Local Ranking*

  In Local Ranking, it can also compute more focused heat normalization term d(i) for people. Thus, Local Ranking could perform better than Global Ranking.

d) Web mining: Web mining - is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining.

B. System Implementation Plan

1) Algorithm Technique

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; γ_{pp}, γ_{ww}, γ_{pw} : thermal conductivity between people, between words, between people and words, respectively.

Output: a ranked list of names according to the query

1 Model Construction
2 Compute the number of distinct co-occurring people *Co(i)* for each person *i* from H_p
3 Construct degree matrices D_p, D_w, D_{ep}, D_{ew} from H_p, H_w and W_e
4 Construct heat normalization matrices $D_{p'}$ by D_p and *Co(i)*'s, and $D_{w'}=D_w$
5 $L_{pp} = \gamma_{pp}D_p^{-1/2}H_pW_eD_{ep}^{-1}H_p^TD_{p'}^{-1} - (\gamma_{pp} + \gamma_{pw})D_p^{1/2}D_{P'}^{-1}$
6 $L_{pw} = \gamma_{pw}D_p^{-1/2}H_pW_eD_{ew}^{-1}H_w^TD_{w'}^{-1}$

$L_{wp} = \gamma_{pw} D_w^{-1/2} H_w W_e D_{ep}^{-1} H_p^T D_{p'}^{-1}$
$L_{ww} = \gamma_{ww} D_w^{-1/2} H_w W_e D_{ew}^{-1} H_w^T D_{w'}^{-1} - (\gamma_{ww} + \gamma_{pw}) D_w^{1/2} D_{w'}^{-1}$
Construct L by L_{pp}, L_{pw}, L_{wp} and L_{ww}
Diffusion and Ranking
for *k=1* to *n* do
  | f = (I+)f
end
Rank people names according to f

Algorithm 2: One Time Re-Ranking

Input: $H_p, H_w, H_e, \gamma_{pp}, \gamma_{ww}, \gamma_{pw}$: as defined in Algorithm 1;
Top: top *k* names after the first run of CoDiffusion;
Scores: corresponding ranking scores of the top *k* names
Output: a ranked list of people names
Initialization query vector f = 0
for i = 1 to *k* do
  | $f_{Top(i)}$ = scores(*i*)
end
Invoke CoDiffusion without global normalization using Parameters $H_p, H_w, W_e, f, \gamma_{pp}, \gamma_{ww}$ and γ_{pw}
Return the ranked list generated by CoDiffusion

Algorithm 3: Iterative Re-Ranking

Input: $H_p, H_w, W_e, \gamma_{pp}, \gamma_{ww}, \gamma_{pw}$: as defined in Algorithm 1; *Top, Scores:* as defined in Algorithm 2; k_0 :deduction of *k* in each iteration; *Iter_num*: number of iteration
Output: a ranked list of people names
for *j=1* to *Iter_num* do
  Initialize query vector f = 0
  for *i = 1* to Length(Top) do
    | $f_{Top(i)}$ = scores(*i*)
  end
  Find pages containing at least two names in *Top* and Construct corresponding H'_p, H'_w and W'_e
  Invoke CoDiffusion without global normalization Using parameters H'_p, H'_w and W'_e, γ_{pp}, γ_{ww}, γ_{pw}, f
  Set *Top* and *Scores* to the top $k - j * k_0$ names and Their corresponding scores outputted by CoDiffusion
end
Return a ranked list according to *Top*

C. System Features

1. Diffusion model defines the experts.
2. Experts provide the quality web pages information.
3. Here we display strong connection query related results as a output content.

III. RESULT

A. Performance Evaluation Result

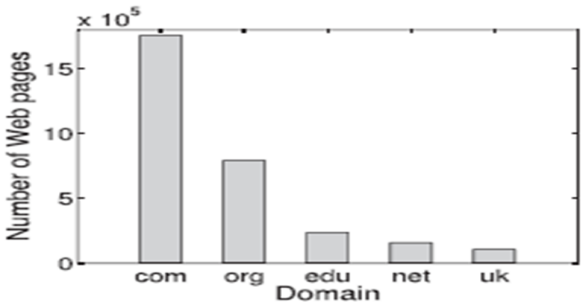

Fig. 3 Top five domains in our data set

IV. CONCLUSION

In this paper we implemented a general expert search problem on the Web. We proposed not to deep parse
Web pages for expert search. Instead, it is possible to leverage co-occurrence relationships such as name-keyword co-occurrences and name-name co-occurrences to rank experts. A ranking algorithm called Co-Diffusion was developed based on this concept. Co-Diffusion adopts a heat diffusion model on heterogeneous hyper-graphs to capture expertise information encoded in these co-occurrence relationships. Experiments on ClueWeb09 and two benchmark datasets consisting of research queries demonstrated that Co-Diffusion outperformed the baseline algorithms significantly. Experiments on conductivity coefficients verified that co-occurrences were indeed useful. We also explored queries other than research related topics and showed that Co-Diffusion could return good results and outperform baselines. Finally, we tried using re-ranking to boost performance.

V. ACKNOWLEDGMENT

With deep sense of gratitude we would like to thanks all the people who have lit my path with their kind guidance. We are grateful to these intellectuals who did their best to help us during our project. It is my proud privilege to express deep sense of gratitude to, Prof. Dr. N. S. Walimbe, Principal of PVG COE, Nashik, for his comments and kind permission to complete this rst phase project. We remain indebted our Prof.M.T.Jagtap HOD of computer Department and our project guide Prof.S.N.Bhadane and project co-ordinator Prof. J.Y.Kapdnis e, of Computer Department for this timely suggestion and valuable guidance. We thank all colleagues for their appreciable help in the project.

REFERENCES

[1] J. 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, “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.
[3] 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. 1012 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 5, MAY 2013 Fig. 9. Running time when varying the number of relevant webpages (Local Ranking).
[4] 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.
[5] 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.
[6] 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.
[7] K. Balog and M. de Rijke, “Combining Candidate and Document Models for Expert Search,” Proc. 17th Text Retrieval Conf. (TREC), 2008.
[8] 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.
[9] 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.
[10] 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.
[11] Ziyu Guan, Gengxin Miao, Russell McLoughlin, Xifeng Yan, Member, IEEE, 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