ISSN 2320 - 2602

Volume 3, No.5, May 2014

International Journal of Advances in Computer Science and Technology

Available Online at http://warse.org/pdfs/2014/ijacst06352014.pdf

Concomitant Supported Dissemination for Expert Search on Web

Gandham Sai[1], Thannir Madhugopal[2]

[1]MGIT,Hyderabad,India, saigandham467@gmail.com

[2]MGIT,Hyderabad,India, thannir.gopal.sachin@gmail.com

ABSTRACT

Expert search has been studied in various contexts. We inspect a commonexpert search problem. Piercing experts on the web, where thousands of names and millions of webpages are considered. It has two challenging issues: 1) web pages could be of full of noises and low quality2) The expertise evidences scattered in web pages are usually arises confusion. We propose to leverage the large amount of concomitant information to assess relevance and reputation of a person name for a query topic. The concomitant structure is modeled using a hyper graph, on which a heat disseminates based ranking algorithm is proposed. Query keywords are regarded as sources of heat and a name of person which has strong link with the query (i.e., frequently concomitant with query keywords and with other names related to query keywords) will receive most of the heat, thus being ranked high. Experiments on the web collection show that our algorithm is effective for retrieving experts and performs baseline algorithms significantly this work would be regarded as one step toward addressing the more common entity search problem without complex NLP techniques.

Key words: Expert search,noises,expertise evidences, concomitant information,hyper graph, NLP techniques

1. INTRODUCTION

We propose anexpert search problem. Expert search on the web[1],[2],[3],[4],[5],[15]. This considers People names and ordinary webpages. This task is different from organizational expert search and is more like Google where our aim is to return a list of experts with good quality. It has some new challenges

1) Compared to an organization's storage, ordinary webpages could be of full of noises and varying quality (e.g., spam).2) Theexpertise evidences spread in webpages are usually ambiguous and vague. In traditional organizational expert search, relevance is the concern. However, considering the challenges mentioned, we also need to consider trustworthiness of data sources and name's reputation for a query topic as well. We suspect the reputation and relevance can be captured by the large amount of name-keyword and name-name concomitants on the web. Using a large amount of concomitant information, noises could be suppressed since noisy co-occurrences would not appear always on the web. A major contribution of this study is an examination of new expert search problem .Searching experts on the web, and the proposal of utilizing concomitant relationships to assess the relevance and reputation of a person name with respect to a query simultaneously. This work would be regarded as one step toward addressing the more general entity search problem without complex NLP techniques, where different types of entities are considered, e.g. locations, people, and organizations. We abstract the concomitant relationships using a heterogeneous hypergraph and develop a novel heat dissemination method on this hypergraph to address the expert search problem. The dissemination method considers both relevance and reputation for ranking experts, as well as the quality of data sources. We also try to improve performance by reranking based on name pseudo relevance feedback.

2. ALGORITHM

The algorithm Concomitant dissemination is shown in Algorithm1.It has two phases: "Model Construction" and "Dissemination and Ranking". In the Model Construction phase, we use the parameters to construct matrix L [9], [13] and given data, which is then used in the Dissemination and Ranking phase to generate the ranked list of people names

Algorithm 2.1: Concomitant dissemination

Input:Hp:weightedincidencematrix between pages and people; Hw:weighted incidence matrix between pages and words;We:diagonal matrix holding PageRank scores of pages;f:the query vector; γpp, γww, γpw: thermal conductivity betweenpeople, between words, between people and words, respectively.

Output: ranked list of names according to the given query

```
1 Model Construction
2Compute the number of distinct co-occurring people Co (i) for each personi from Hp
3Construct degree matricesDw, Dp, Dep, Dew by Hp, We,Hw,
4Construct heat normalization matrices Dp′ by Dp and Co(i)'s, andDw′ = Dw
5Lpp= γppD−1/2p HpWeD−1ep HTp D−1p′ − (γpp + γpw)D1/2p D−1p′
6Lpw=γpwD−1/2pHpWeD−1ewHTwD−1w′
7Lwp=γpwD−1/2wHwWeD−1epHTp D−1p′
8Lww=γwwD−1/2wHwWeD−1ewHTwD−1w′ − (γww + γpw)D1/2wD−1w′
```

```
Construct L by Lpp, Lpw, Lwp and Lww
Dissemination and Ranking
for k = 1 to n do
    f = (I + Ln)f
end
Rank people names according to f
```

Algorithm 2.2: One-Time Re-Ranking

Input:Hp,Hw,We,γpp,γww,γpw:as defined in Algorithm1;T op: top k names after the first run of CoDiffusion[1]; Scores:Corresponding rank scores of the top k names

Output: a ranked list of people names

```
Initialize query vector f = 0
for i = 1 to k do
    fTop(i) = Scores(i)
end
Invoke CoDiffusion without global normalization using parametersHp, Hw, We, f, γpp, γww and γpw
Return the ranked list generated by CoDiffusion[1]
```

Algorithm 2.3: Iterative Re-Ranking

Input:Hp,Hw,We,γpp,γww,γpw: as defined in Algorithm1;T op,Scores:asdefined in Algorithm2;k0:deduction of k ineach iteration; Iternum: number of iterations

Output: a ranked list of people names

```
for j = 1 to Iternum do
    Initialize query vector f = 0
    for i = 1 to Length (T op) do
        fTop (i) = Scores (i)
    end
    Find pages containing at least two names in T op and construct corresponding H′p, H′w and W′e
    Invoke CoDiffusion without global normalization using parameters H′p, H′w, W′e, f, γpp, γww and γpw
    Set T op and Scores to the top k − j ∗ k0 names and their corresponding scores outputted by CoDiffusion
Return a ranked list according to T op
```

Experts are retrieved based on the rankings implemented by the above algorithms.

3.MODULES

In order to implement concomitant supported dissemination the system has been identified to have following modules.

1. Creation of forum and implement the search engine
2. Heat Dissemination on Heterogeneous Hyper graphs
3. Dissemination Model
4. Expert Search
5. Global ranking versus Local ranking
6. Performance evolution results.

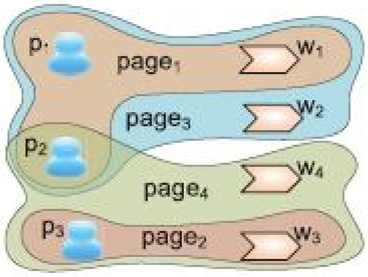
3.1 Creation of forum and implement the search engine

An Internet forum, or message board, is an online site where people can exchange information in the form of posted messages. A discussion forum is tree-like or hierarchical in structure.A forum can hold a number of sub forums, each of which may have many topics. Within a forum's topic, each new discussion started is a thread, and can be replied to by as many people as so wish.

Renlifang [1], [7], [5] is an object level search engine which allows users to query about people, locations, organizations and explore their relationships. It employs relation extraction techniques and entity extraction. The major technique used in search engines like Renlifang is to extract structural information about entities and their relationships by deep-parsing web pages.

3.2 Heat Dissemination on Heterogeneous Hyper graphs

In a hyper graph, each edge (called hyper edge) can connect two or more vertices. Formally, let G (V, E) be a hyper graph with vertex set V and edge set E. A hyper edge e € E can be regarded as a subset of vertices. e is said to be incident with a vertex v if v € e. Each hyper edge e is associated with a weight denoted by w(e). In our problem setting[14], 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, we can construct a heterogeneous hyper graph [1], [11], [12] 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 e's 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.



3.3 Dissemination Model

Heat diffuses[10] in a medium from positions with higher temperatures to those with lower temperatures. The most important property of heat dissemination 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 dissemination model is constructed as follows: At time t, each vertex i € V will receive an amount of heat from its neighbors.

3.4 Expert Search

An 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[6] on the web, e.g., a professor would co-occur with professors and with many students; a senior forum user would actively answer questions for other users and consequently concomitant with many different users.

3.5 Global ranking versus Local ranking

There are two possible schemes to implement our algorithm: 1) we perform "Model Construction" [1], [8] on the entire web collection and for each query we only need to perform the "Dissemination and Ranking" part in Algorithm 1.The first phase of Algorithm 1 needs to be done only once. After, the constructed model is used for queries. We call this scheme as Global Ranking. 2) We first obtain related web pages for a query by querying the web collection.

4.CONCLUSION

In this project, we studied a common expert search problem on the web. We proposed for expert search by not to deep-parse webpages. Instead, it is possible to leverage concomitant relationships such as name-keyword co-occurrences and name-name co-occurrences to rank experts. A ranking algorithm called Concomitant-dissemination was developed based on this concept. CoDissemination adopts a heat diffusion model on heterogeneous hyergraphs to capture expertise information encoded in these co-occurrence relationships. Experiments on two benchmark data sets consisting of research queries demonstrated that CoDissemination outperformed the baseline algorithms significantly. Research on conductivity coefficients resulted that cooccurrences were indeed useful.

REFERENCES

1. Guan,Ziyu,Miao,Gengxin,McLoughlin,Russel,Yan, Xifeg,Cai,Deng"**Co-occurrence based diffusion for expert search on web**" vol 25,pp1001-1014,2013.
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.
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. R. Bekkerman and A. McCallum, "**Disambiguating web Appearances of People in a Social Network,**" Proc. Int'l Conf. World Wide Web (WWW), pp. 463-470, 2005.
12. M. Belkin and P. Niyogi, "**LaplacianEigenmaps for Dimensionality Reduction and Data Representation,**" Neural Computation, vol. 15, no. 6, pp. 1373-1396, 2003.
13. P.R. Carlile, "**Working Knowledge: How Organizations Manage What They Know,**" Human Resource Planning, vol. 21, no. 4, pp. 58- 60, 1998.
14. N. Craswell, A.P. de Vries, and I. Soboroff, "**Overview of the Trec 2005 Enterprise Track,**" Proc. Text Retrieval Conf. (TREC), 2005.
15. N. Craswell, D. Hawking, A.M. Vercoustre, and P. Wilkins, "**Panoptic Expert: Searching for Experts not Just for Documents,**"Proc. Ausweb Poster, 2001