Query Based Expert Search Based on Relevance Class and Web Page Quality Ranking

Francy.J.,(M.E)¹
Computer Science and Engineering,
Kathir College of Engineering,Coimbatore.

Anandhi.R.,(M.E)²
Computer Science and Engineering,
Kathir College of Engineering,Coimbatore.

Aiswarya.R.,(M.E)³
Computer Science and Engineering,
Kathir College of Engineering,Coimbatore.

Sharmila.V.,(M.E)⁴
Computer Science and Engineering,
Kathir College of Engineering,Coimbatore.

Shabariram.C.P.,(M.E)⁵
Computer Science and Engineering,
Kathir College of Engineering,Coimbatore.

Abstract-Expert search is mostly used in the areas of academic groups, organizations. The general expert search problem which is observed is searching experts on the web where lot of web pages and experts names are considered. It has two demanding issues namely: 1) The quality and noisy content of web pages. 2) The names of the experts are usually unclear. We propose to influence the huge amount of coexisting information to access relevance and reputation of a person name for a query. The coexisting structure is modeled using a hypergraph, on which a heat dissipation based ranking algorithm is proposed. The keywords in the query are concerned as sources of heat, and name of the person which is closely related with the query will receive most of the heat, and ranked high. Our algorithms are efficient for retrieving experts and surpass baseline algorithms notably. This work would be noticed as one step toward addressing the problem in searching experts from the web.

Index Terms-Expert search, dissipation, relevance, coexisting

1 INTRODUCTION

Industry and academic groups turned its attention towards expert search. The enterprise track at TREC has as primary goal to build a test collection for enterprise search. The expert search problem was addressed in other domains such as academic groups, internet forums and question answering. However, early work on expert search is mostly used in the areas of enterprises, academic groups and internet forums. Recently, the need to find experts on a variety of daily life topics is increasing. Normally, users search for people who can answer their natural language questions by using the system which requires users to register and join the community. Whereas, the web contains large amount of information about people(e.g., personal
home pages, blogs). Therefore, it is possible to build a powerful expert search engine by manipulating the information about the people on the web.

In this paper, we propose a general expert search problem which is observed is searching experts on the web where lot of web pages and experts names are considered. This problem is different from enterprise expert search and is more like Google where our goal is to return a list of experts with reasonable equitable quality. It has demanding issues: 1) Each web page could be of varying quality and full of noises. Examples of noises from a news page of CNN, i.e., links to famous news stories and commercial informations. 2) The name of the experts scattered in web pages are usually unclear. Wikipedia page of Ana Ivanovic, a former World No.1 tennis player could be the example. The web page consists of information that she used to practice in a swimming pool, even though she is not an expert in swimming.

In traditional enterprise expert search, relevance is the major burden. However, considering the demanding issues specified above, we also need to consider a name’s reputation for a query topic as well as the trustworthiness of data sources. We speculate the relevance and reputation can be captured by the large amount of keyword-name and name-name co-existence on the web. Noises in the web page could be suppressed using a huge amount of co-existence information since noisy co-existence would not appear regularly on the web. The problem in the second example can thus be alleviated because Ana Ivanovic apparently does not co-exist frequently with prominent swimmers. In precise, we intend to direct the new demanding issues by influencing the linkage of experts sported on the web: 1) Relevance. Relevant experts should co-exist frequently on many web pages with the keywords given in the query. 2) Reputation. Relevant experts should co-exist frequently with other people related to the query provided by the user, irrespective of whether they are experts or not. For example, a notable 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-exist with many other users. 3) Trustworthiness. Relevant experts tend to occur in high quality webpages.

Since nowadays peoples are socialized and social activities shall be reflected on the web the second observation could be true for many domains. Following these observations, we propose to model the co-exist relationships among people names and words in a heterogeneous hypergraph where webpages are treated as hyperedges with PageRank scores as their weights. Then, we renovate a unique heat dissipation model on the hyper graph. Grounded on this model, an expert ranking algorithm, called Co-existence Dissipation is developed. Given a query, we consider keywords in the query as heat sources and perform heat dissipation. Names with highest heat sources are returned. Instinctively, people who have close connection with query will be ranked high. Intrinsically, CoDissipation accumulates data collected from various webpages. This accumulation could be a
good remedy for noises in web data. CoDissipation complements traditional language model based methods, if it applies their relevance scores.

Relation with Renlifang: Renlifang is an object level search engine which allows users to query about people, locations, and organizations and discover their relationships. It employs entity abstraction and relation abstraction 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. In contrast, CoDissipation does not depend on complicated natural language processing techniques to search experts. The ranking algorithm used by Renlifang for searching the experts is not publicly known.

Our contributions: A major contribution of this study is an examination of a new expert search problem: searching experts on the web, and the proposal of utilizing co-existing relationships to access the relevance and reputation of a person name with respect to a query simultaneously. This work would be noticed as one step toward addressing the problem in searching experts on the web, where different types of entities are considered, e.g., people, place, organizations. We abstract the co-existing relationships using a heterogeneous hypergraph and develop a novel heat dissipation method on this hypergraph to address the expert search problem. The dissipation method considers both relevance and reputation for providing rank to the experts, as well as the quality of data sources. In addition to the above specification, we try to boost performance by reranking based on name pseudo relevance feedback. Our algorithms are efficient for retrieving experts and surpass baseline algorithms and well-known language model-based approaches significantly. We also exhibit the purpose of people co-existing information in ranking experts.

2 RELATED WORK

Expert search is a growing research area. Early approaches of expert search involve building a knowledge base which contains the descriptions of people’s skills within an organization. However, manual work of creating a knowledge base is time consuming and lengthy. Hence, automatic approaches have been developed for building people profiles. Expert search become a hot research area since the start of the TREC enterprise track in 2005. A lot of studies were emerged to organizational expert search. In the area of enterprise search, Balog et al. proposed a language model framework for expert search [3]. The model 1 is equivalent to profile centric approach where text from all the documents associated with a person is amassed to denote that person. The model 2 is a document centric approach 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 person. This process was formulated in a generative probabilistic model. Balog et al. exposed that model 2 performs good than model 1 [3] and it become one of the most prominent methods for expert search. In their future work, Balog
et al. tried to employ and process the language model on a smaller data set comprising multilingual data crawled from Tilburg University’s website [8].

Researchers have investigated using additional information to boost retrieval performance, such as page rank, in degree, and URL length of documents, person-person similarity [8], internal document structures that indicate people’s association with document content [6], query expansion and relevance feedback using people names [5], nonlocal evidence [4], proximity between existence of query keywords and people names [2]. Besides language models, other methods have been proposed for organizational expert search. Macdonald and Ounis proposed a model based on voting and data fusion techniques [10]. Serdyukov et al. modeled associations between people and documents as a bipartite graph and performed probabilistic random walks to find relevant experts. Fang et al. proposed a relevance-based discriminative learning framework for expert search. Many other methods for organizational expert search were proposed during TREC enterprise track.

Two benchmark data sets, provided for the organizational expert search works are W3C and CSIRO, which are crawls of the websites of World Wide Web Consortium and Commonwealth Scientific and Industrial Research Organization, respectively. Ordinary web pages and people names makes organizational expert search to become different from searching experts on the web. The major interest in organizational expert search is relevance, whereas we also need to focus the reputation of the person. This is because 1) compared to an organization’s website or document repository, web collections of could be of low quality and noisy; 2) the expertise information contained in ordinary web pages could be vague. In this paper, we propose to use co-existences to access the relevance and reputation of a person name with respect to a query simultaneously and we will demonstrate its effectiveness in experiments.

There are other expert retrieval problems. Balog and de Rijki studied the problem of finding similar experts, given example experts [7]. Zhang et al. studied characteristics of internet forums and tested using link analysis methods to identify users with high expertise. Liu et al. studied expert finding in community based question answering websites and treated it as an IR problem [9]. Mimno and McCallum used topic modeling to address the problem of matching papers with reviewers. Later Karimzadehgan et al. addressed this review assignment problem based on matching of multiple aspects of expertise. Deng et al. explored using language modeling and a topic based model for expert finding in the DBLP bibliography data. Zhou et al. proposed co-ranking authors and their publications using coupled random walks.

Finally, our work is also related to heat dissipation on graphs. In real world, heat dissipates in a medium from position with a higher temperature to that with a lower temperature. The idea of heat dissipation was extended to the discrete graph model, with applications such as dimension reduction, classification, antispamming,
social network marketing and online advertisement matching [1]. These studied considered dissipation in homogeneous graphs. In this paper, we develop a dissipation model based on heterogeneous hypergraphs for our expert search problem.

3 PROBLEM FORMULATION

In a hypergraph, each edge (called hyperedge) can connect two or more vertices. Formally, let \( G=(V,E) \) be a hypergraph with vertex set \( V \) and edge set \( E \). A hyperedge \( e \in E \) can be regarded as a subset of vertices. \( e \) is said to be incident with a vertex \( v \) if \( v \in e \). 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 web pages, denoted by \( P \), \( W \), and \( D \), respectively.

By the co-existing relationships among \( P \) and \( W \) established by web pages, we can construct a heterogeneous hypergraph \( G_{P,W}=(V,E) \) where \( V \) contains vertices representing all the people and words and each \( e \in E \) corresponds to a web page. 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.

We propose using heat dissipation to address this ranking problem. Let \( V_p \) and \( V_w \) represent the vertex set corresponding to people and words, respectively. Consequently, \( V = V_p \cup V_w \). Let \( H_p \) be a \( |V_p| \times |E| \) weighted incidence matrix where an entry \( H_p(v,e) = wt_{v,e} \) if \( v \in e \) (\( v \in V_p \)) and 0 otherwise. \( H_w \) is defined similarly for \( V_w \). \( wt_{v,e} \) reflects the connection strength between object \( v \) and page \( e \). We set \( H_p(v,e) \) to the number of times person \( v \) appears in page \( e \) and set \( H_w(v,e) \) to the TF-IDF score of word \( v \) in \( e \).

The degree of a vertex \( v \) is defined as

\[
d(v) = \left\{ \begin{array}{ll}
\sum_{e \in E} w(e) H_p(v,e) & v \in V_p \\
\sum_{e \in E} w(e) H_w(v,e) & v \in V_w
\end{array} \right.
\]

The degree of a hyperedge \( e \) is defined as

\[
\delta(e) = \delta_p(e) + \delta_w(e),
\]

where \( \delta_p(e) = \sum_{v \in V_p} H_p(v,e) \) and \( \delta_w(e) = \sum_{v \in V_w} H_w(v,e) \). We define \( f^P_i(t) \) and \( f^W_i(t) \) to be the heat of vertex \( i \in V_p \) and that of vertex \( j \in V_w \) at time \( t \), respectively. Let \( f^P(t) \) and \( f^W(t) \) be the heat distribution vectors at time \( t \) with sizes \( |V_p| \times 1 \) and \( |V_w| \times 1 \), 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 as heat sources and rank their objects according to the heat distribution at time \( t \), which reflects the affinity between the objects and heat sources. This is a general form of ranking model. In our problem formulation, words are queries and we are required to get the ranking of people.

4 DISSIPATION MODEL

Heat dissipates in a medium from positions with higher temperatures to those with lower temperatures. The most prominent property of heat dissipation is that the flow of heat at a point is directly proportional to the second order derivative of heat with respect to the space at the particular point. In this work,
we impose our dissipation model on heterogeneous hypergraphs. Different medium suppose to have different thermal conductivity coefficients. Hence, we define three coefficients namely $\gamma_{pp}$, $\gamma_{ww}$, and $\gamma_{pw}$ to describe the thermal conductivity in the midst of people, amongst words, and between people and words, respectively. The dissipation model is constructed in the form of matrix denoted as $L$, which has the block structure given as below,

$$L = \begin{bmatrix} L_{pp} & L_{pw} \\ L_{wp} & L_{ww} \end{bmatrix}$$

5 INTERPRETATION OF THE MODEL

The perception behind the dissipation model is by constructing the matrix $L$, we essentially accumulate the co-existing information among people and words to exhibit the bondage strength between each pair of objects. This accumulation could be useful for sorting with noises on the web pages. After the dissipation matrix construction we distribute heat from query keywords on this accumulated structure. Importantly, names having close connection with both query keywords and related names and words will be ranked high.

6 DISSIPATION FOR EXPERT SEARCH

The algorithm co-dissipation consists of two phases namely “model construction” and “dissipation and ranking of experts”. In the phase of model construction, we implement the given data and parameter to construct the matrix $L$, then it is applied to dissipation and ranking phase to generate the ranked list of people names. The ranking phase can be done either with Global Ranking or Local Ranking. In global ranking, the model construction on the entire web collection is performed and for each query we only need to perform the dissipation and ranking. In local ranking, the related web pages for the query are collected and the model is constructed then the dissipation is applied. Compared to global ranking, local ranking could perform better since it focus on relevant web pages. Here, we suggest reranking algorithms to improve the ranking results in order to improve prominent names for the query. The ranking result which is concluded will not depart too much from the original one.

7 CONCLUSIONS

In this paper, we examined a general expert search problem on the web. We suggest not to deep-parse web pages for expert search alternatively, it is possible to use co-existing relationships such as name-keyword co-existences and name-name co-existences to rank experts. Based on this concept a ranking algorithm called co-dissipation was developed. The expertise information embedded in the co-existence relationships is retrieved by adopting heat dissipation model on heterogeneous hypergraphs. The
co-dissipation algorithm performs better than baseline algorithms significantly. Co-dissipation outperforms baseline in queries other than research related topics.

REFERENCES


