International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878
Volume 2 Issue 4, April 2014

A Global Fusion Re-Ranking for Effective Data Search with Correlation Model

A. Sudha[1]

[1]PG scholar, Computer Science and Engineering, United Institute of Technology, Tamil Nadu, India,

Abstract: ***Web service plays an important role in e-business and e-commerce applications. The web service applications are interoperable and it will work on any platform, large scale distributed systems can be developed easily. Surfing on the internet is becoming more prominent in day today life. In searching there are major issues like noisy data and unwanted data. The existing system provides additional results rather than appropriate results and uses PageRank algorithm. PageRank uses link analysis algorithm to measure the page relevance in a hyperlinked set of documents. In order to improve the existing co–diffusion of keywords and ranking, the system introduces the result remerging and re-ranking concepts. Basically ranking will be performed by the popularity, key term and its frequency count. In the proposed system an enhanced ranking concept is used which improves the performance of the co-diffusion ranking system. The ranking is done locally for two or more search engines and a global re-ranking are done at the end considering the page structure which includes pre-link, post link and the popularity as well. All these are performed by using the newly proposed Semantic Dual Correlation Algorithm which makes searching effective.***

Keywords: Web mining, Expert search, performance, data mining, Expert Search Information Retrieval.

1. Introduction

The Web is a distributed hypermedia system or a system where the responsibility for creating content is distributed among many people. Web browsers display a Web document and enable users to link to other Web pages. Web servers respond to the requests of browsers. They find and send requested resources back to the browser. The following are the ways to find information on the Web (1) Browse or surf the Web – This involves linking from one Web page to another, and then (2) Search the Web – This method involves using search engines to locate Web pages with the information that we are looking for. Sites are ranked based on their importance. Horizontal bars displayed next to each web page indicate the importance of the page. There are different types of searches:

- Index searches (Directory)
- Natural Languages Searches
- Concept /keyword Searches

1.1 Index Searches

A Directory of catalogued, hierarchically, structured lists of web sites, like the yellow pages of a phone directory are example for index search. Limited web site inclusion and it is not good for complex or specific concept/keyword searches. Typically these are evaluated for popularity, content and/or quality before being included.

1.2 Natural Language Searches

It is a directory of possible questions. Good for frequently asked questions, use on simple questions, least comprehensive in sites included in search. Developers have developed general question structures that can be asked about a topic. A question is typed into the question box; possible alternative statements of the question are then given, followed by links for possible answers.

1.3 Concept/Keyword Searches

A concept search (or conceptual search) is an automated information retrieval method that is used to search electronically stored unstructured text (for example, digital archives, email, scientific literature, etc.) for information that is conceptually similar to the information provided in a search query. In other words, the ideas expressed in the information retrieved in response to a concept search query are relevant to the ideas contained in the text of the query. Most organize the sites in terms of relevance as determined by links to a site and/or actual frequency of access of a web page, “hits”. Concepts/keywords are typed into the query box using Boolean logic and the search engine’s rules to limit the list to the most relevant.

1.4 Web Mining

Web mining is an application of data mining techniques to discover patterns from the web. Web mining is divided into three types;

a) **Web Usage Mining**
Web usage mining is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking on the Internet. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications.

b) **Web Content Mining**
Web content mining is the process of extracting useful information from contents of web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of

text, images, audio, video, or structured records such as lists and tables.

c) **Web Structure Mining**
Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds:

- Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.
- Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

The rest of this paper is organized as follows. Section II gives a brief background on related work. Section III introduces our proposed system approach. Section IV discusses the experiments and results. Section V Finally concludes the paper and outlines future research avenues.

2. Background and Related Work

A distributed search engine framework, in which every web server answers queries over its own data, Results from multiple web servers will be merged to generate a ranked hyperlink list on the submitting server. This paper presents a series of algorithms that compute Page Rank in such framework.[1] The preliminary experiments on a real data set demonstrate that the system achieves comparable accuracy on Page Rank vectors to Google's well known Page Rank algorithm and, therefore, high quality of query results. Internet search engines, such as Google, use web crawlers to download data from the Web. The crawled data is stored on centralized servers. For instance, Google computes Page Rank to evaluate the importance of pages. Thus, crawled web data repository has two impacts on the results of a query. First, more qualified results may be found in a larger data set. Second, more web pages will provide a bigger link graph which, in turn, will result in a more accurate Page Rank computation.[1] Internet search engines have popularized the keyword based search paradigm. While traditional database management systems offer powerful query languages, they do not allow keyword-based search.[2] Existing techniques use either documents (as a whole) or proximity-based techniques to represent candidate experts. Proximity-based techniques have shown clear precision enhancing benefits.[3] An alternative to keyword search is structured search where users direct their search by browsing classification hierarchies. Both models are tremendously valuable success of both keyword search and the classification hierarchy are evident today. [2] Searching for information about people in the web is one of the most common activities of many internet users.[4] While such structured searches over databases are no doubt useful, unlike the documents world, there is little support for keyword search over databases.[2] By introducing the TREC in 2005 finding experts is getting more interest within the research community. Numerous models have been proposed that rank candidates by their level of expertise with respect to some topic. A component is used to estimates the strength of the association between a document and a person. [2] Around 30% of search engine queries include person names. Retrieving information about people from web search engines can become difficult when a person has nick names or name aliases.[4] For instance, in a normal search engine, it may be desirable to produce a diverse ranking of documents for ambiguous queries, to satisfy more possible distinct user needs. The future research directions from this work are two-fold: Firstly, it is clear that the problem of topic drift does occur, particularly within the expert search task. Further measures that can show when and how topic drift is occurring during QE would be beneficial. Secondly, the successful application of QE to expert search introduces other potential applications, such as finding similar experts, creating a diverse ranking of candidates for ambiguous queries, and even the automatic creation of a 'roadmap of expertise' in an organization.[6].Generative models such as statistical language modelling have been widely studied in the task of expert search to model the relationship between experts and their expertise indicated in supporting documents.[5] On the other hand, discriminative models have received little attention in expert search research, although they have been shown to outperform generative models in many other information retrieval and machine learning applications.[5]

3. Proposed System

In order to enhance the existing co–diffusion of keywords and ranking, the proposed work introduces the result remerging and re-ranking concepts. Basically ranking will be performed by the popularity, key term and its frequency count. By using following technique an enhanced ranking concept is used which improves the performance of the co-diffusion ranking by considering some additional parameters.

3.1 Proposed Flow Diagram

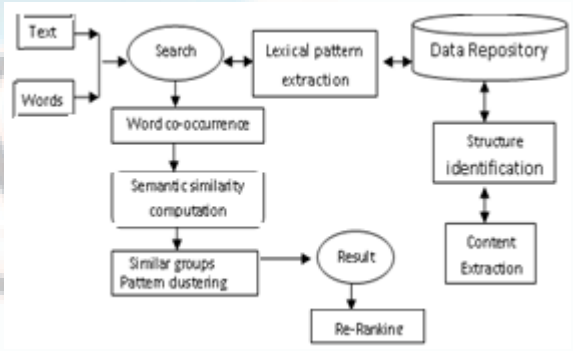

Figure 1: Architecture diagram

The system proposes the use co-occurrences to assess the relevance and reputation of a person name with respect to a query simultaneously. The system introduces a diffusion model based on heterogeneous hyper graphs for the expert search problem.

International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878
Volume 2 Issue 4, April 2014

In this paper, we have studied a general expert search problem on the web. The system proposed not to deep-parse WebPages for expert search. Instead, it is possible to leverage co-occurrence relationships such as follows, 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.

3.2 Enhancements

- A fusion model which combines and re-rank the results from two different search engines.
- The proposed search engine is not domain based; this is a general search engine.

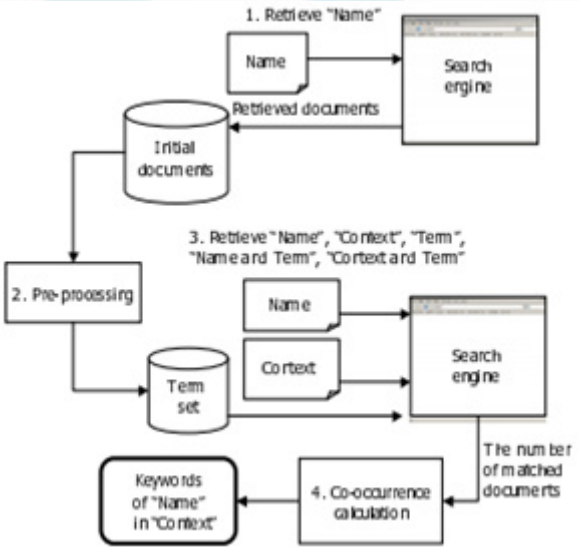

Figure 2: Keyword extraction

4. Experiments and Results

4.1 Page Rank Vector

It is a general algorithm and the idea of page rank is the importance of any web page can be judged by looking at the pages link to it. If we create a web page 'j' and include a hyperlink to the webpage 'j', here 'j' is considered as important and relevant for the topic. Here a lot of page link to 'j' so 'j' is important. If 'j' has only one back link, but becomes from an authoritative site 'k' like yahoo, Google, and CNN etc. So 'k' transfers its authority to 'j' so that 'k' asserts that 'j' is important. So we consider the web net as directed graph with nodes as web pages and edges as links between them.

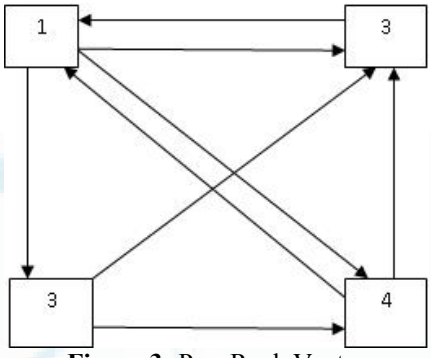

Figure 3: PageRank Vector

Here in this model each page transfers evenly its importance to the pages that it links to. Node 1 has 3 outgoing edges, so it will pass on 1/3, and so on2, 3 and 4. The incoming links are considered for writing the values, and the matrix is written with no. of incoming nodes.

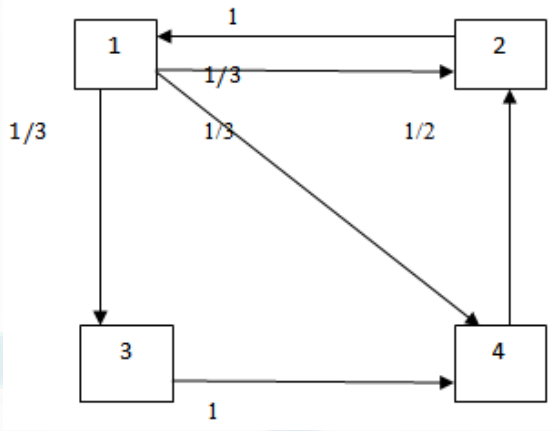

Figure 4: Page Rank Vector Ranking

Let A be the transition matrix of the graph, A

$$A=\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 0 & 1 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 0 & 1 & 0 \end{bmatrix}$$

First the importance is uniformly distributed among 4 nodes and it is denoted as ¼(.25).The initial rank vector has entries equal to ¼. The incoming links increase the importance of the web page. The rank of each page is added by the current value of the incoming links. This is done by multiplying the vector 'v' by matrix 'A'. The importance vector is denoted as v,

Step 1: A new vector is calculated Av,

$$\begin{bmatrix} .25 \\ .25 \\ .25 \\ .25 \end{bmatrix}$$

International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878
Volume 2 Issue 4, April 2014

$$Av=\begin{pmatrix}0 & 1 & 0 & 0\\ 1/3 & 0 & 0 & 1\\ 1/3 & 0 & 0 & 0\\ 1/3 & 0 & 1 & 0\end{pmatrix}\begin{pmatrix}.25\\ .25\\ .25\\ .25\end{pmatrix}$$

$$Av=\begin{pmatrix}.25\\ .33\\ .08\\ .33\end{pmatrix}$$

Step 2: A new vector is calculated A^2v,

$$A(Av)=\begin{pmatrix}0 & 1 & 0 & 0\\ 1/3 & 0 & 0 & 1\\ 1/3 & 0 & 0 & 0\\ 1/3 & 0 & 1 & 0\end{pmatrix}\begin{pmatrix}.25\\ .25\\ .25\\ .25\end{pmatrix}$$

$$A^2v=\begin{pmatrix}.25\\ .44\\ .02\\ .44\end{pmatrix}$$

The iterations goes upto A^kv. The sequence of iterates are Av,A^2v,…… A^kv. And the below value is for A^6v,

$$v^*=\begin{pmatrix}.25\\ 1.36\\ .00\\ 1.36\end{pmatrix}$$

This is called as PageRank vector of the given web graph. Here the nodes 2 and 4 has the highest number of hits and then comes the node 1 and then the node 3.

5. Conclusion

In this paper it is implemented for avoiding the noisy or unwanted data's. In the existing they have used Page Rank, it is used as link analysis algorithm to measure the page relevance in a hyperlinked set of documents. Basically ranking will be performed by the popularity, key term and its frequency count. In this paper an enhanced ranking concept is used which improves the performance of the co-diffusion ranking by considering some additional parameters. The additional parameters are page structure, post link and pre link are considered and the ranking is done locally for two or more search engines and a global re-ranking are done at the end. In order to enhance the existing co–diffusion of keywords and ranking, the system introduces the result remerging and re-ranking concepts. By this the noisy data's are avoided and an effective search has been done.

References

[1] Yuan Wang and David J. DeWitt, "*Computing Page Rank in a Distributed Internet Search System*", Proceedings of the 30th VLDB Conference, 2004
[2] 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.
[3] 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.
[4] Bharathi and Sumana Manickan "*Lexical Pattern- Based Approach for Extracting Name Aliases*" in Springer, 2008.
[5] Yi Fang, Luo Si and Aditya P. Mathur "*Discriminative Models of Integrating Document Evidence and Document-Candidate Associations for Expert Search*", *ACM, 2010.*
[6] 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.

Author Profile

A. Sudha is currently pursuing a Master of Engineering in Computer Science and Engineering at United Institute of Technology, affiliated to Anna University Chennai.