Middle-East Journal of Scientific Research 23 (7): 1414-1419, 2015
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2015.23.07.22322

Relevant Information Retrieval Using Relation Based Semantic Page Ranking Algorithm

[1]Veeramalai Sankaradass, [2]K. Vijayabhaskar, [1]P. Meenakshi and [3]A. Kannan

[1]Department of Computer Science and Engineering, Vel. Tech. High Tech. Dr. Rangarajan Dr. Sakunthala Engineering College, Avadi, Chennai, India
[2]Department of Computer Science and Engineering, Misrimal Navajee Munoth Jain Engg College, Thorapakkam, Chennai, India
[3]Department of Information Science and Technology, Anna University, Chennai, India

Abstract: Relevance is highly important concept in information retrieval (IR), but it is hard to define. Retrieval results, indexing, etc., are evaluated with methods that are based on the concept of relevance. Retrieval of relevant information and personalization is a challenging one and essential for the users who are all looking for relevant information. Here, a recommendation system for Relevant Information Retrieval has been proposed and implemented using Relation Based Semantic Page Ranking Algorithm. The proposed page ranking algorithm supports relation based matching and approximate structure matching in the semantic web. The effectiveness of the proposed algorithm is shown by comparing with existing algorithm. The precision and recall of the retrieved information have been analyzed and the improvement of the proposed relation based page ranking algorithm is shown. The proposed new page ranking algorithm which combines both key and related documents is used to rank the pages retrieved web pages to improve the relevancy of the recommendation system.

Key words: Information retrieval · Personalization · Relation based page · Ontology · Page ranking

INTRODUCTION

As the Internet usage rate is rapidly increasing, the volume of electronic documents present in the web that matches the user's interest can be seen on the web has also substantially increased. Currently, the user gives a set of keywords as input to a search engine and the search engine returns a list of pages that are related to the keyword topics or terms. Since the result page set consists of too much irrelevant information, the relevant information required for a user are retrieved only after a few keyword modifications. In particular, users have to browse and view sites one after another for a long time until they are satisfied to have a good set that is relevant to their interest or more likely, they give up the search. Moreover, the users want to have effective search techniques to find the relevant information easily and precisely from the web. However, most of the existing search engines provide information most of which may not be relevant to the user queries.

Moreover, common search scans return a vast number of web pages, many of which will be irrelevant to the searcher and of which only about ten or twenty of the top ranked results are normally browsed

Ranking is the key to success for any search engine. Its popularity remains with its ability to order search results in a meaningful way for users, so that those documents most useful to them are placed high up on the list. The primary objective of ranking is to reduce the burden on users having to sift through lots of irrelevant information in order to find the information most suited to their request.

The conventional approach to ranking documents is based on simple metrics such as:

- Counting the number of times the query terms appear in the document,
- Considering if they appear in the title or at the start of the document,

Corresponding Author: Veeramalai Sankaradass, Department of Computer Science and Engineering, Vel. Tech. High Tech. Dr. Rangarajan Dr. Sakunthala Engineering College, Avadi, Chennai, India.

- Their structural and display features such as bold and italics.

These approaches are not only very simple but their usefulness in determining relevancy is questionable. Additionally, they analyze each document in complete isolation, can only be used to rank documents that contain the specific query terms and ignore any notion of topical importance.

Key Documents Ranking: In this method, ranking is based on an intelligent and highly sophisticated algorithm that orders individual documents by considering not just the document itself, but all the other documents within its semantic neighborhood (Those documents that are semantically related).

Related Documents Ranking: A dynamic ranking approach for ordering the Related Documents list is used in this approach. The rank of a document here depends on its semantic similarity to the topics within the key documents list. This ensures that they are organized with the most semantically similar documents to the user's needs at the top.

Here, a new page ranking algorithm which combines both key and related documents is used to rank the pages retrieved to improve the relevancy of the recommendation system.

Related Work: Recently, modern computer network and internet technologies connect people distributed at different places in the world and ease the delivery of information. Thereupon, users are enabled to share the knowledge with other people using the web technologies. However, it becomes very tedious for a web surfer to browse the connecting web pages one by one due to the large extent of the unstructured web. Thus search engines have been adopted as a solution to overcome such problems over the past few years. In addition to the search engines, web mining plays a leading role in providing relevant information to the users since it can speed up the exchange of knowledge hidden in volatile collections of data on the Internet. The proposed recommendation system retrieves and ranks the relevant web pages to the user desires.

There are many keyword-based search engines that are available for Web search and these search engines often return a long list of search results, many of which are not what the user wants. Thus, users may spend lots of time on running through all links of the list to find the truly relevant information.

One way to reduce the time spending on browsing search results is to provide personalized Web search. Through personalized web searching, web pages relevant to user interests are ranked in the front of the result list, thus leading to a quick process for the users to just access. While the amount of web pages is growing at a rapid speed, the issue of devising a personalized Web search is of increasing importance [1].

Ranking the search results is a fundamental problem in information retrieval. Most common approaches focus primarily on the similarity of a query and a page, as well as the overall page quality [2,3]. PageRank is a link analysis algorithm to measure the page relevance in a hyperlinked set of documents, such as the World Wide Web. This algorithm assigns a numerical weight to each document. This numerical weight is also called PageRank of the document. Diligent *et al.* [3] proposed a general probabilistic framework for Web page scoring systems, which incorporates and extends many of the relevant models proposed in the literature. The PageRank of a web page represents the likelihood that a person randomly clicking will arrive at this page. Maratea *et al.* [4-17] explained that the PageRank algorithm requires several iterations to be executed. For each iteration, the values are better approximated to the real value.

The purpose of Page Ranking is to measure the relative importance of the pages in the web. There are many algorithms for this purpose. The most important ones are: Hyper Search, Hyperlink-Induced Topic Search (HITS), Page Rank and Trust Rank.

Hyper Search Algorithm: Hyper Search has been the first published technique to measure the importance of the pages in the web. This algorithm served as a base for the next ones [5].

Hyperlink-Induced Topic Search: Two popular webpage ranking algorithms are HITS and PageRank. The HITS (Hyperlink Induced Topic Selection) algorithm is developed by Kleinberg makes the distinction between hubs and authorities and compute them in a mutually reinforcing way [6]. This algorithm gives more weightage to inlinks and less weight age to out link.

Page Rank: In 1998, Page and Brin developed the link-based ranking algorithm called PageRank [7]. PageRank calculates the authoritativeness of web pages based on a graph constructed by web pages and their hyperlinks, without considering the topic of each page. Since then, much research has been explored to differentiate authorities of different topics. PageRank was

developed by Google and is named after Larry Page, Google's co-founder and president. PageRank ranks pages based on web structure [18,8]. It measures the importance of the pages by analyzing the links. This is a commonly used algorithm in web structure mining.

Trust Rank: Trust Rank is a semi-automatic link analysis technique for separating useful WebPages from spam. The HITS and the PageRank algorithms can also be used for this purpose, but they have been subject to manipulation. To avoid this manipulation, the Trust Rank algorithm selects a small set of documents. These sets of documents are evaluated by an expert. Once the reputable seed pages are manually identified, a crawl extending outward from the seed set seeks out similarly reliable and trustworthy pages.

Works on Page Rank Ing Algorithms: The work of Brian et al. [9] provides empirical analyze of the belief that linked pages tend to be related and analyzed the hypothesis that an anchor text provides a succinct summary of a target page. Haveliwala [10] proposed Topic-sensitive PageRank, which performs multiple PageRank calculations, one for each topic. When computing the PageRank score for each category, the random surfer jumps to a page in that category at random rather than just any web page. This has the effect of biasing the PageRank to that topic. This approach needs a set of pages that are accurately classified. Moreover, Wenpu *et al.* [11] described that the PageRank also correlates very closely to in degree.

Eirinaki *et al.* [12] presented a hybrid probabilistic predictive model extending the properties of Markov models by incorporating link analysis methods. More specifically, they proposed the use of a PageRank-style algorithm for assigning prior probabilities to the web pages based on their importance in the web site's graph. Hitoshi Nakakubo *et al.* [13] proposed link act technique. This technique defines the link act, "Act of recommending linking ahead". The score obtained by this technique helps to shows the level from which each Web page is referred. However, link act is not necessarily possible compared with web pages, which wants to be recommended.

Nie *et al.* [14] proposed another web ranking algorithm that considers the topics of web pages. In that work, the contribution that each category has to the authority of web pages is distinguished by means of soft classification, in which a probability distribution is given for a web page being in each category. Kohlschutter et al [15] conducted analysis on expanding these topic categories and showed that ranking performance increases with the topic level up to a certain point.

Ja-Hwung Su [16] proposed a novel approach for personalized page ranking and recommendation by integrating association mining and Page Rank so as to meet user's search goals. Lamberti *et al*. [17] studied a relation-based page rank algorithm to be used in conjunction with Semantic Web search engines that simply relies on information that could be extracted from user queries and on annotated resources. However, the web pages retrieved by applying these techniques are not relevant to the users in many occasions. Therefore, in this paper, semantic based approach for relevant information has been proposed.

In this work, a new page ranking algorithm is proposed in which the following metrics are considered. It used the number of times the query terms appear in the page, title or at the start of the pages. Moreover, the user's ratings are also considered to rank the pages. The web page ranking process helps to improve the relevancy of the system.

System Architecture: To retrieve relevant information, in this work, ontologies are used to integrate user knowledge and rule schemas are used to filter the uninteresting rules. Furthermore, an interactive framework is designed to assist the user throughout the analyzing task. Moreover, the quality of the filtered rules was validated by the domain expert at various points in the interactive process. Semantic analysis techniques are used in this work that coordinates with word frequency counts to extract relevant information with page ranking techniques.

Figure 1 shows the relevant web page retrieval process. In this system, the rule base allows formalizing user knowledge and goals. Domain expert offers a general view over user knowledge in database domain and user expectations express the prior user knowledge over the discovered rules.

The rule manager is acting in major rule. The rule manager will choose the suitable rules generated by the system based on the different conditions of system and users with different possibilities. Additionally, the rule manager will refer the domain expert as well as domain ontology based on that the relevant web page will be retrieved.

First of all, heterogeneous Web pages are collected. And the collected web pages are undergoing for the semantic annotation with respect to ontology schema.

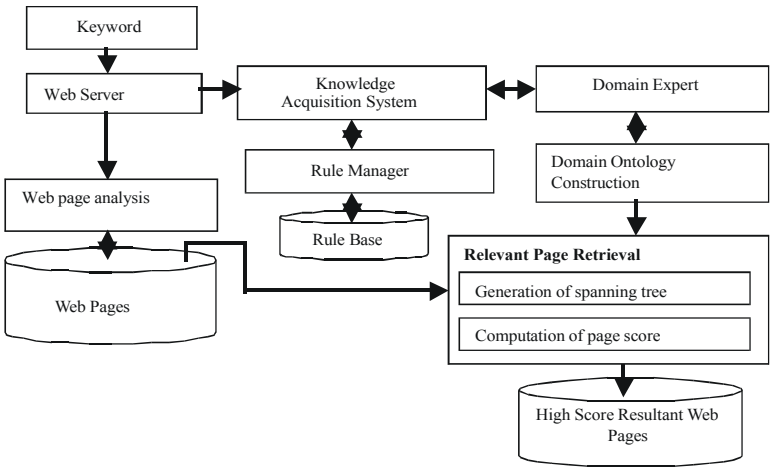

Fig. 1: Proposed System Architecture

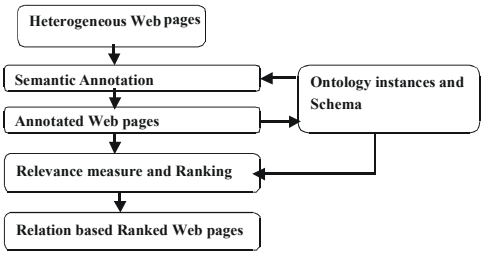

Fig. 2: Flow diagram for relation based ranking

The annotated web pages retrieved and the relevancy of the retrieved web pages are measured and ranked the retrieved web pages with relation concept.

In the semantic page analysis, according to the concept associated with keyword, the annotated web pages are collected from semantic web. The collected web pages are analyzed with ontology concept. And finally the computer based spanning tree graph was generated. From this spanning tree, the high accuracy of the large page relevancy was measured based on the edges.

Ranking is the key to success for any search engine. It lives or dies on its ability to order search results in a meaningful way for users, so that those documents most useful to them are placed high up on the list. The primary objective of ranking is to reduce the burden on users having to sift through lots of irrelevant information in order to find the information most suited to their request.

Proposed Algorithm

Relation Based Semantic Page Ranking Algorithm:

Step 1: Specify the user keyword

Step 2: Collect annotated web pages from semantic web.

Step 3: Create Knowledge base with collected web pages

Step 4: Analyze the concept associated with concept using ontology

Step 5: Construct the initial result set from each page using rules

Step 6: Compute the page sub graph

Step 7: Compute page scores by generating spanning tress

Step 8: Compute the final result

Step 8: To calculate relevance score

- Generate all possible page trees
- Check whether page spanning tree is obtained
- Obtain higher accuracy of page relevance by having larger number of edges

In all the way, the normal procedures will be followed without deviating the data mining concept. Finally, the retrieved pages are ranked based on semantic relation and provides higher relevancy of retrieved documents.

RESULTS AND DISCUSSION

Precision and recall are the basic measures used in evaluating search strategies. Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.

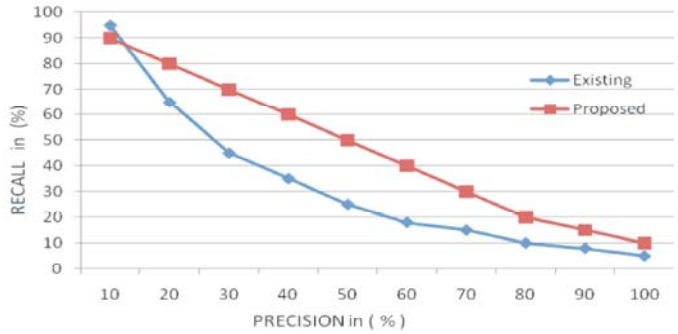

Fig. 3: Relationship between Precision and Recall

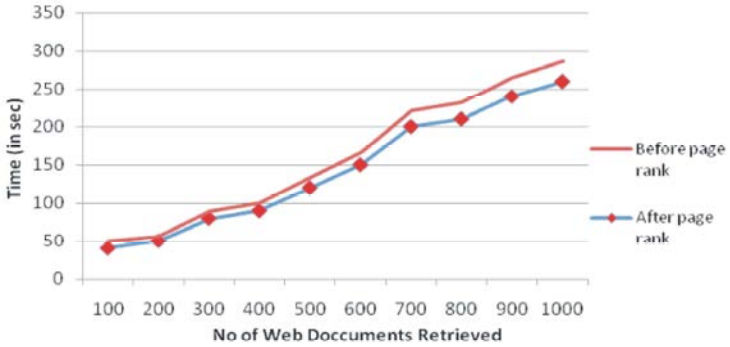

Fig. 4: Web Document Retrieval Analyses with Respect to Time

Table 1: Results of Session #1

Exp	Precision	Recall
1	77.52	92.14
2	77.14	95.46
3	78.18	93.42
4	78.65	88.23
5	79.51	94.14
6	79.87	93.56
7	78.37	93.65

Table 2: Comparison of the proposed algorithm with existing algorithm

Algorithm	Page Ranking	Weighted Page Ranking	Relation Based Page Ranking
Mining technique use	WSM	WSM	WSM,WCM &concept using Ontology
Working	Computes scores at Indexing time. Results are sorted according to importance of Pages.	Computes scores at Indexing time. Results are sorted according to Page importance	Compute page scores by generating spanning tress. Higher accuracy of page relevance by having larger number of edges
I/P Parameter	Back links	Back links, Forward links	Spanning Tree
Relevancy	Medium	High	Higher
Search Engine	Google	Research model	Research model

Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage. In the graph shown in Figure 3, the two lines represent the performance of the proposed and existing system. While the exact slope of the curve may vary between the existing and proposed systems, the general inverse relationship between recall and precision remains.

In Figure 4, the analysis of page ranking algorithm is shown. The relevancy measures obtained before and after the proposed page ranking techniques are compared here with respect to time.

Table 1 shows the relationships and performance improvements in precision and recall with different experiments for the session 1. The same way we have analyzed precision and recall values for different session.

In the above table, the existing ranking algorithm with respect to page for relevant information retrieval (Rekha Jain and Purohit G.N 2011) is compared with proposed algorithm. In the proposed algorithm, the mining techniques which use ontology and the spanning tree algorithm.

CONCLISION

In this paper, new algorithms have been proposed and implemented for relevant web page retrieval. The relevancy of the proposed relation based semantic page ranking algorithm is compared with existing algorithm. The proposed algorithm gives 10% improvement. The relationship between precision and recall analyzed and improved to the sufficient level to rank the retrieved web pages to the user interest.

REFERENCES

1. Wen-Chih Peng and Yu-Chin Lin, 2006. Ranking Web Search Results from Personalized Perspective, Proceedings of the 8th IEEE International Conference on E-Commerce Technology and the 3rd IEEE International Conference on Enterprise Computing, E-Commerce and E-Services, pp: 12-17.
2. Salton, G. and M.J. Mcgill, 1983. Introduction to Modern Information Retrieval", McGraw Hill Publishers, New York.
3. Diligenti, M., M. Gori and M. Maggini, A Unified Probabilistic Framework for Web Page Scoring Systems, 2004. IEEE Transactions on Knowledge and Data Engineering, 16(1): 4-16.
4. Maratea, A. and A. Petrosino, 2009. An Heuristic Approach to Page Recommendation in Web Usage Mining, Ninth International Conference on Intelligent Systems Design and Applications, pp: 1043-1048.
5. Marchiori Massimo, 1997. The Quest for Correct Information on the Web: Hyper Search Engines, World Wide Web Conference Series Computer Networks and ISDN Systems, 29(8-13): 1225-1236.
6. Kleinberg, J.M., 1999. Authoritative Sources in A Hyperlinked Environment, Journal of ACM, 48: 604-632.
7. Lawrence Page and Sergey Brin, 1998. The Anatomy of a Large-Scale Hyper textual Web Search Engine, Journal of Computer Networks and ISDN Systems, 30(1-7): 107-117.
8. White, R.W., J. Jose and I. Ruthven, 2003. A Task-Oriented Study on the Influencing Effects of Query-Biased Summarization in Web Searching, Information Processing and Management, 39(5): 707-733.
9. Davison, Brian D., 2000. Topical Locality in the Web', Proceedings of the Twenty Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp: 272-279.
10. Haveliwala, T., A. Gionis, D. Klein and P. Indyk, 2002. Evaluating Strategies for Similarity Search on the Web, In Proceedings of the Eleventh International World Wide Web Conference, pp: 432-442.
11. Wenpu Xing and Ali Ghorbani, 2004. Weighted PageRank Algorithm, In Proceedings of the Second Annual Conference on Communication Networks and Services Research, pp: 305-314.
12. Eirinaki, M. and M. Vazirgiannis, 2005. Usage-Based Pagerank for Web Personalization, In Proceedings of 5th IEEE International Conference on Data Mining, pp: 130-137.
13. Hitoshi Nakakubo, Shinsuke Nakajima, Kenji Hatano, Jun Miyazaki and Shunsuke Uemura, 2007. Web Page Scoring Based on Link Analysis of Web Page Sets", In Proceedings of DEXA Workshops, pp: 269-273.
14. Nie, L., B.D. Davison and X. Qi, 2006. Topical Link Analysis for Web Search", In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval, ACM Press, pp: 91-98.
15. Kohlschutter, C., P.A. Chirita and W. Nejdl, 2007. Utility Analysis for Topically Biased Pagerank, In Proceedings of the 16th International Conference on World Wide Web, ACM Press, pp: 1211-1212.
16. Ja-Hwung Su, Bo-Wen Wang and V.S. Tseng, 2008. Effective Ranking and Recommendation on Webpage Retrieval by Integrating Association Mining and Page Rank, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 3: 455-458.
17. Lamberti, F., A. Sanna and C. Demartini, 2009. A Relation-Based Page Rank Algorithm for Semantic Web Search Engines, IEEE Transactions on Knowledge and Data Engineering, 21(1): 123-136.
18. Rekha, Jain and G.N. Purohit, 2011. Page Ranking Algorithms for Web Mining, International Journal of Computer Applications (0975 - 8887), 13(5).