Semantic Relation Based-Ranking Approach for Web Search Engines

Y.K.Viswanadham, K.K.Praneeth Tellakula, I.Ananda Babu, Kalamkuri Srikanth

***Abstract*—The majority of Semantic Web search engines retrieve information by focusing on the use of concepts and relations restricted to the query provided by the user. In this paper, we propose 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. Relevance is measured as the probability that a retrieved resource actually contains those relations whose existence was assumed by the user at the time of query definition.**

***Index Terms*—Web Search, Ontology, Ranking, Concepts.**

I. INTRODUCTION

With the tremendous growth of information available to end users through the Web [5], search engines come to play ever a more critical role. Nevertheless, because of their general-purpose approach, it is always less uncommon that obtained result sets provide a burden of useless pages. The next-generation Web architecture, represented by the Semantic Web [2], [5] provides the layered architecture possibly allowing overcoming this limitation. Several search engines have been proposed, which allow increasing information retrieval accuracy by exploiting a key content of Semantic Web resources, that is, relations. However, in order to rank results [3], most of the existing solutions need to work on the whole annotated knowledge base.

The Semantic Web is trying to close the gap between user demand and the need for hyperlink accessibility. This approach deals with two issues: (1) common formats for integration and combination of data drawn from diverse sources, as opposed to the original Web which mainly focused on the interchange of documents; and (2) the language for recording how the data relates to real world objects.

These two features allow a person, or a machine, to start off in one database and then move through an unending set of databases, which are not connected by wires but connected by topic. This information networking is based on the idea of semantic associations, where one entity (node) is connected to another entity (node) by means of a relationship (an edge).

Most search engines retrieve information accurately by exploiting key content of associations in Semantic Web resources, or relations. We propose a relation-based page rank algorithm to be used in conjunction with Semantic Web search engines which relies on information that could be extracted from user queries and the ontology for a given page.

Y.K.Viswanadham, Department of Information Technology, Gudlavalleru Engineering College, Gudlavalleru, India,9948717476

K.K.Praneeth Tellakula, Department of Information Technology, Gudlavalleru Engineering College, Gudlavalleru, India,9885053541

I.Ananda Babu, Department of Information Technology, Gudlavalleru Engineering College, Gudlavalleru, India,9533531959

Kalamkuri Srikanth, Department of Information Technology, Gudlavalleru Engineering College, Gudlavalleru, India,9848748844

Relevance score is measured as the probability that a given resource contains those relations which existed in the user's mind at the time of query definition. The idea is to use existent relations in the ontology, named "virtual links" and apply them to a set of pages to increase the probabilities of finding the implicit relations made by the user at the time of the query.

In this paper, we will prove that relations among concepts [1] embedded into semantic annotations can be effectively exploited to define a ranking strategy for Semantic web search engines. This sort of ranking behaves at an inner level that is, it exploits more precise information that can be made available within a web page and can be used in conjunction with other established ranking strategies to further improve the accuracy of query results. With respect to other ranking strategies for the Semantic web, my approach only relies on the knowledge [6] of the user query, the web pages to be ranked, and the underlying ontology. Thus, it allows us to effectively manage the search space and to reduce the complexity associated with the ranking task.

II. PREVIOUS WORK

The idea of exploiting ontology-based annotations for information is not new; semantic search engine would consider keyword concept associations and would return a page only if keywords (or synonyms, homonyms, etc.) are found within the page and related to associated concepts. The success is measured by the "predictability" that the user would have guessed such an association exists.

In the semantic model proposed in [13], a ranking system is created based on an estimate of the probability that keywords and/or concepts within an annotated page "A" are linked to one another in a way that is the same or similar to the one in the user's mind at the time of query definition.

A. Anatomy of search engine

The web creates new challenges for information retrieval. The amount of information on the web [5] is growing rapidly, as well as the number of new users inexperienced in the art of web research [7]. People are likely to surf the web using its link graph, often starting with high quality human maintained indices such as Yahoo or with search engines. Human maintained lists cover popular topics effectively but are subjective, expensive to build and maintain, slow to improve, and cannot cover all esoteric topics. Automated search engines that rely on keyword [9] matching usually return too many low quality matches. To make matters worse, some advertisers attempt to gain people's attention by taking measures meant to mislead automated search engines.

B. System Features

The search engine has two important features that help it produce high precision results. First, it makes use of the link structure of the Web [5] to calculate a quality ranking for each web page. This ranking is called Page Rank [3]. Second, it utilizes link to improve search results.

C. *Page Rank: Bringing Order to the Web*

The citation (link) graph of the web is an important resource that has largely gone unused in existing web search engines. We have created maps containing as many as 518 million of these hyperlinks, a significant sample of the total. These maps allow rapid calculation of a web page's "Page Rank" [6], [3], an objective measure of its citation importance that corresponds well with people's subjective idea of importance. Because of this correspondence, Page Rank is an excellent way to prioritize [9] the results of web keyword searches. For most popular subjects, a simple text-matching search that is restricted to web page titles performs admirably when Page Rank prioritizes the results.

D. *Description of Page Rank Calculation:*

Academic citation literature has been applied to the web, largely by counting citations or back links [4] to a given page. This gives some approximation of a page's importance or quality. Page Rank extends this idea by not counting links from all pages equally, and by normalizing by the number of links on a page. Page Rank is defined as follows:

we assume page A has pages T1...Tn, which point to it (i.e., are citations). The parameter d is a damping factor, which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also C(A) is defined as the number of links going out of page A [3]. The Page Rank of a page A is given as follows:

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

Note that the Page Ranks form a probability distribution over web pages, so the sum of all web pages' Page Ranks will be one.

Page Rank or PR(A) can be calculated using a simple iterative algorithm [2], and corresponds to the principal eigenvector of the normalized link matrix of the web. Also, a Page Rank for 26 million web pages can be computed in a few hours on a medium size workstation.

III. Proposed System

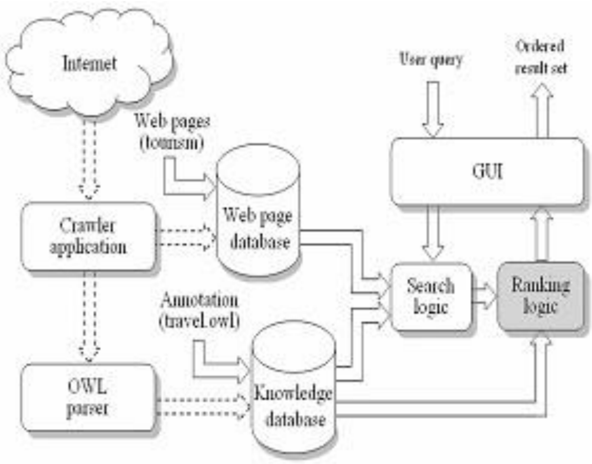

Fig 1:Proposed System for Web Search Engine model

A. *Crawler Application*

This program will run as a multithreaded program continuously, which will take the link [4] and download that page, it will extract links form the downloaded page [7] and again will download those pages this will repeat continuously. The downloaded pages are saved in the Web pages database.

B. *OWL Parser*

This program will retrieve the downloaded pages and for each page it will remove HTML [9] tags as well as any special characters. The pages will now contain only the data. The data is saved in the Knowledge database.

C. *Web page Database*

It is used for storing downloaded HTML pages.

D. *Knowledge Database*

It is used for storing the page which only has data and all the HTML and special characters [10] are removed.

E. *GUI*

GUI is the program which will take input from the user and display the output to the user. The input is the keyword which the user sends as the query and the output will be web pages returned which will have higher ranking.

F. *Search Logic*

This program will retrieve the pages from the database [3] and will check for the keyword that the user has entered. It will only retrieve the pages that matched the keyword [5]. Then program will consider the retrieved pages and construct the sub graph [8].

G. *Ranking Logic*

This program will consider all the pages that are retrieve as well as the sub graphs and will compute page spanning forests [2]. Using this it will compute the scores based on the relations. And merge it with the original web pages saved in the web [5] page database. The ranking is calculated based on the scores of the concept words. The result is out to the user.

To evaluate the feasibility of this new method, a controlled Semantic Web environment was constructed. To do this, we must generate controlled ontologies and page subgraphs, and then modify its relations in order to make it more suitable for demonstrating the method’s functionality. The architecture workflow will look like this:

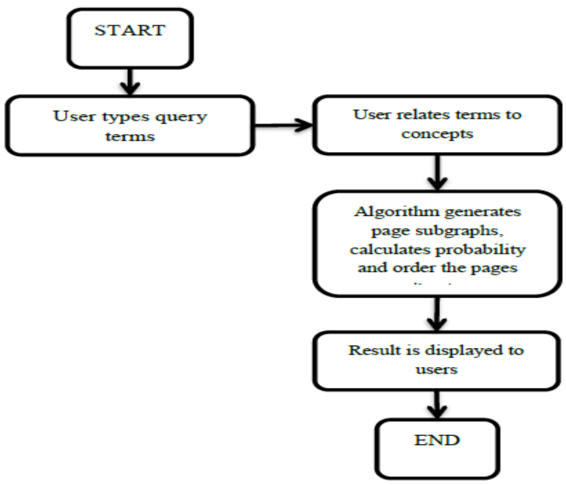

Fig 2: Architecture for workflow

IV. Results

The objective of this experimentation is to measure the contribution of the inclusion of semantics in the ranking of results returned by search engines. The idea is to display results according to ranking generated by our system scheduling results according to the ontology driven approach that we propose, we refer to this ranking by 'semantic ranking'.
The concept [3] of this paper is implemented and different results are shown below

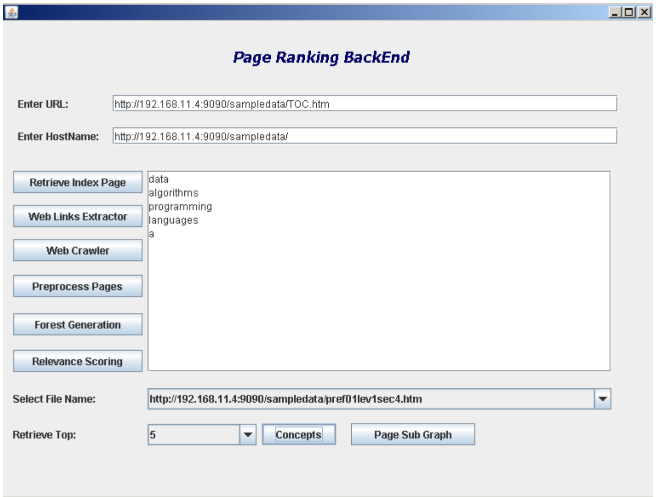

Fig 3:GUI for Web Search Engine

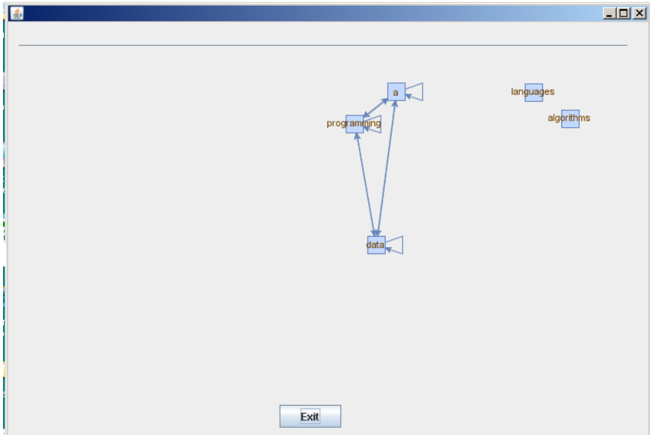

Fig 4:Sub Graphs

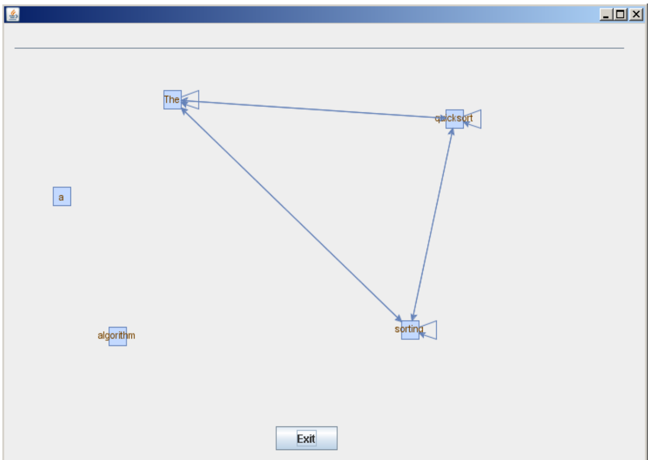

Fig 5:Sub Graphs

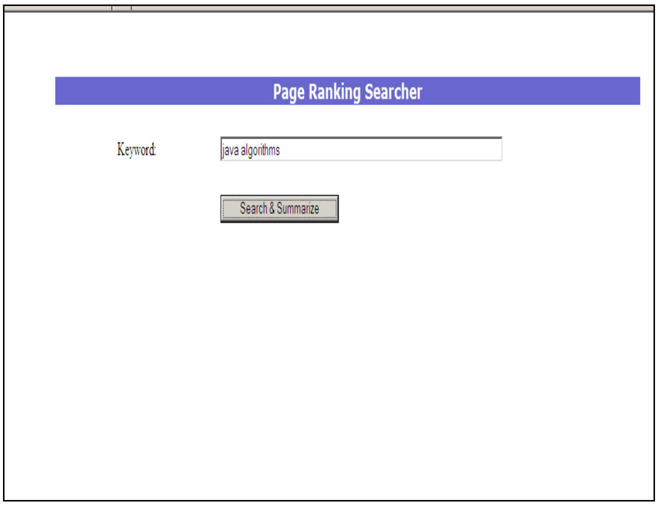

Fig 6: Page Ranking Searcher

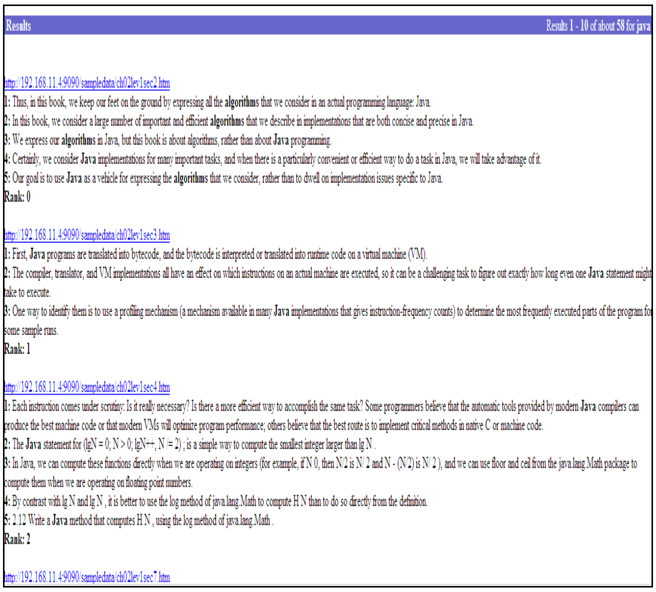

Fig 7: Result Page

V. Performance Analysis

The proposed paper is implemented in Java and Servlets technology on a Pentium-Dual Core PC with 320 GB hard-disk and 1G RAM with apache web server. The propose paper's concepts [3] shows efficient results and has been efficiently tested on different Messages.

VI. Conclusion

The next-generation Web architecture represented by the Semantic Web will provide adequate instruments for improving search strategies and enhance the probability of seeing the user query [2] satisfied without requiring tiresome manual refinement. However, actual methods for ranking the returned result set will have to be adjusted to fully exploit additional contents characterized by semantic [1] annotations including ontology-based concepts and relations. Several ranking algorithms for the Semantic Web [5] exploiting relation-based metadata have been proposed. Nevertheless, they mainly use page relevance criteria based on information that has to be derived from the whole knowledge base, making their application often unfeasible in huge semantic environments.

In this work, we propose a novel ranking strategy that is capable of providing relevance score [7] for a Web page into an annotated result set by simply considering the user query [4], the page annotation, and the underlying ontology. Page relevance is measured through a probability-aware approach that relies on several graph-based representations of the involved entities. By neglecting the contribution of the remaining annotated resources, a reduction in the cost of the query answering phase could be expected. Despite the promising results in terms of both time complexity and accuracy, further efforts will be requested to foster scalability into future Semantic Web repositories based on multiple ontologies, characterized by billions of pages, and possibly altered through next generation "semantic" [1] spam techniques.

REFERENCES

[1] B. Aleman-Meza, C. Halaschek, I. Arpinar, and A. Sheth, "A Context-Aware Semantic Association Ranking," Proc. First Int'l Workshop Semantic Web and Databases (SWDB '03), pp. 33-50, 2003.

[2] K. Anyanwu, A. Maduko, and A. Sheth, "SemRank: Ranking Complex Relation Search Results on the Semantic Web," Proc. 14th Int'l Conf. World Wide Web (WWW '05), pp. 117-127, 2005.

[3] R. Guha, R. McCool, and E. Miller, "Semantic Search," Proc. 12th Int'l Conf. World Wide Web (WWW '03), pp. 700-709, 2003.

[4] Y. Lei, V. Uren, and E. Motta, "SemSearch: A Search Engine for the Semantic Web," Proc. 15th Int'l Conf. Managing Knowledge in a World of Networks (EKAW '06), pp. 238-245, 2006.

[5] Y. Li, Y. Wang, and X. Huang, "A Relation-Based Search Engine in Semantic Web," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 2, pp. 273-282, Feb. 2007.

[6] A. Pisharody and H.E. Michel, "Search Engine Technique Using Keyword Relations," Proc. Int'l Conf. Artificial Intelligence (ICAI '05), pp. 300-306, 2005.

[7] T. Priebe, C. Schlager, and G. Pernul, "A Search Engine for RDF Metadata," Proc. 15th Int'l Workshop Database and Expert Systems Applications (DEXA '04), pp. 168-172, 2004.

[8] A. Sheth, C. Bertram, D. Avant, B. Hammond, K. Kochut, and Y. Warke, "Managing Semantic Content for the Web," IEEE Internet Computing, pp. 80-87, 2002.

[9] R. Sun, H. Cui, K. Li, M.Y. Kan, and T.S. Chua, "Dependency Relation Matching for AnsIr Selection," Proc. ACM SIGIR '05, pp. 651-652, 2005.

[10] T. Tran, P. Cimiano, S. Rudolph, and R. Studer, "Ontology-Based Interpretation of Keywords for Semantic Search," Proc. Sixth Int'l Semantic Web Conf., pp. 523-536, 2007.

[11] Web Ontology Language, http://www.w3.org/2004/OWL/, 2004.

[12] H. Yang, I. King, and M.R. Lyu, "DiffusionRank: A Possweble Penicillin for Web Spamming," Proc. ACM SIGIR '07, pp. 431-438, 2007.

[13] Fabrizio, L., A. Sanna, and C. Demartini. "A Realtion-Based Page Rank Algorithm for Semantic Web Search Engines." *IEE Transactions on Knowledge and Data Engineering* 21.1 (2009): 123-36.

Y.K.Viswanadham is working as Associate Professor at Gudlavalleru Engineering College, GUDLAVALLERU, AP, INDIA. He is doing Ph.D in S.V.University,TIRUPATHI. His research area Cloud Security and Semantic Web.

K.K.Praneeth Tellakula is working as Assistant Professor at Gudlavalleru Engineering College, GUDLAVALLERU, AP, INDIA. His main research interest includes Mobile networks, Internet of Things, Cloud Computing, and Semantic Web

I. Ananda Babu is working as Assistant Professor at Gudlavalleru Engineering College, GUDLAVALLERU, AP, INDIA. His main research interest includes Cloud Computing, and Distributed Systems

Kalamkuri Srikanth is working as Assistant Professor at Gudlavalleru Engineering College, GUDLAVALLERU, AP, INDIA. His main research interest includes Cloud Computing, Semantic web, Internet of Things and Compiler Constructions.