A Study of Different Ranking Approaches for Semantic Search

Darshan Bhansali
Student, Department of Computer Engineering.
Dwarkadas J. Sanghvi College of Engineering

Harsh Desai
Student, Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering

Khushali Deulkar
Assistant Professor,Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering

ABSTRACT
Search Engines have become an integral part of our day to day life. Our reliance on search engines increases with every passing day. With the amount of data available on Internet increasing exponentially, it becomes important to develop new methods and tools that help to return results relevant to the queries and reduce the time spent on searching. The results should be diverse but at the same time should return results focused on the queries asked. Relation Based Page Rank [4] algorithms are considered to be the next frontier in improvement of Semantic Web Search. The probability of finding relevance in the search results as posited by the user while entering the query is used to measure the relevance. However, its application is limited by the complexity of determining relation between the terms and assigning explicit meaning to each term. Trust Rank is one of the most widely used ranking algorithms for semantic web search. Few other ranking algorithms like HITS algorithm, PageRank algorithm are also used for Semantic Web Searching. In this paper, we will provide a comparison of few ranking approaches.

General Terms
RaRe [3], SemRank [2], WWW, HITS.

Keywords
Rational Research, Semantic Search, Hybrid Spreading Activation [1], Semantic Web.

1. INTRODUCTION
Many of the technological advances achieved in the past decade are either a direct or indirect product of World Wide Web. World Wide Web aka WWW has been at the root of heights of advancement attained by humans. However, the constant increase in the amount of information available on WWW has given rise to unforeseen problems of great complexity. Traditional Information retrieval methods treat documents as single entity and consider them to be similar. Also traditional IR methods are plagued by problems of speed and tend to come up with a lot of insignificant and irrelevant results. The probability that a user will encounter irrelevant search results has steadily increased over time. Thus the need to develop an alternative method for performing search became essential. Semantic Web Search bypasses the boundaries of traditional IR methods by trying to understand not only the contextual meaning of the query but also takes into consideration the assumptions the user makes while typing that query. Semantic Web Search enables search engines to be able to produce relevant search results in a more efficient manner. For this purpose it maintains a library of all the metadata known as a knowledge database. [8]

2. SEMANTIC WEB AND SEMANTIC SEARCH: OVERVIEW
It is important to understand the difference between Semantic Web and Semantic Web Search in order to elucidate various types of ranking approaches employed.

Semantic Web refers to a set of technologies that are used for range of operations performed on the data varying from storing of data, representing it in appropriate format and performing querying operations on data. One of the basic objectives of Semantic Web is to extract data from files of various formats and sources. Unlike the World Wide Web, Semantic Web does not treat a single web page as a page but rather works to find the meaning of each tiny detail and at the end pieces it together to return it as a cohesive result to the query.

Semantic Web Search is concerned with producing relevant results that are based on the research of relationship between the search results and meanings of the keyword outside the domain of its contextual meaning. It implies that Semantic Search comes up with better search results based on relations and by following a procedure of disambiguation until it attains a state of maximum plausibility. It returns search results based on various factors other than the occurrence of exact keyword. It takes into consideration the assumption the user makes while firing the query and tries to establish a relation between the content of various search results. This leads to reduced time spent in searching and better efficiency. [6][10]

2.1 Components of Semantic Search
Semantic Web Search performs basic operations of crawling, Indexing, Linguistic Post-processing or Disambiguation and Searching. A representation of the architecture of Semantic Search is given in Figure 1.

2.1.1 Source File Archive
Source file archives consist of a knowledge database and an inference engine. Knowledge database contain vast amount of data and is aware of the facts of world. The knowledge is not viewed as a procedural code rather the facts are reasoned. The inference engine is responsible for verification of these facts and deduces new facts.

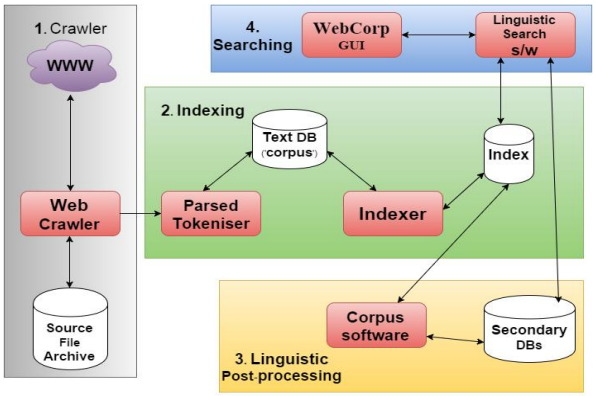

Figure 1: Semantic Web Search Engine Architecture

2.1.2 Web Crawler

The large repository of data available on WWW makes it a laborious task to navigate the repository for indexing. Web Crawlers are responsible for performing this task by following a systematic approach. Beginning with a list of URLs as seeds it visits the URLs, identifies the hyperlinks and adds them to a list and visits them recursively based on a set of policies. Combination of selection, revisit, politeness and parallelization policies determine the behavior of a Web Crawler.

2.1.3 Linguistic Search

Unlike other search engines which solely rely on statistical and analytical algorithms, Semantic Search employs techniques involving linguistic science for the determination of semantic relationship between the keywords of a query. Linguistic search softwares try to gain a holistic view by expanding their domain outside the contextual meaning of the keyword in queries and return results that are relevant to the keywords and help in reducing the search time.

3. DIFFERENT APPROACHES FOR SEMANTIC SEARCH

There exist various approaches for ranking of results in Semantic Search. SHOE, PageRank, HITS are few of the approaches that have been around for a decade now. Each of them is unique and has its own characteristics and advantages. We have given an overview of few different approaches that we have compared in section 4 in this section. [7][9]

3.1 SHOE

The SHOE[5] approach uses semantic markup language to properly describe the context of web pages in order to improve the efficiency of the results which are returned using search engines.

3.1.1 Introduction

The SHOE search implementation involves user to specify the context of his query and SHOE search uses this context to help the user to build a query by example. SHOE uses annotation, which is a process used to add markup to web pages. The SHOE is based on domain-ontology where the document types relate to ontological concepts. For e.g. School home page, Staff's homepage, Staff's Homepage etc., it may contain properties like Teacher's name or the subject they teach. These different web pages relate themselves to ontological concepts and different types of properties using the SHOE markup-language which is not known by the browsers but by only to Semantic based search engines.

3.1.2 Working

With the help of Semantic Search the user is allowed to select a concept and specify properties from the ontology and then the system returns results that are related to the selected concepts including its properties. Considering the example of a School's Website, the user wants to search or select a concept "teacher's homepage" and specifies value from the name property section as "Ram", then the SHOE approach will return teacher's homepages that belongs to teachers with the name "Ram". The main requirement of SHOE is strong coupling between the concepts which the user is interested in and the web-pages.

3.2 Hybrid Spreading Activation

In Hybrid Spreading Activation, approach a combination of traditional search algorithms are used along with spread search techniques for any given semantic model.

3.2.1 Introduction

Hybrid approach is more efficient while searching for queries which consist of keywords that are ambiguous in nature. Spread search technique is widely used for information retrieval applications for Semantic Networks. Whenever an initial set of concepts and its corresponding initial activation values are given, spread search technique is used to find the related concepts in the given ontology. They are basically used to measure the strength of the relation. Hybrid approach consists of a mixture of weight mapping techniques and spread activation technique.

3.2.2 Working

In Weight mapping a numerical value is assigned to each relation instance in the network. However, there exists no known formula that can prove any particular solution to be the best but different measures have been tested in order to create an optimized formula to determine the strength of the relation in the knowledge base.

Spread Activation Techniques employs graphs, whose links are assigned a label based on the ontological definitions. The links are further assigned a numerical weight based on the weight mapping techniques. Taking a set of concepts for reference, corresponding set of closely related concepts are found by traversing through the graph. Few nodes are treated as initial node and have an initial activation value. As the graph is traversed, nodes are activated and a set of nodes is obtained. Distinct weights can be assigned to distinct nodes. Thus each node has an initial value assigned and are placed in a priority queue and the one with the maximum weight is taken out of the queue for processing.

3.3 SemRank

The ranking of search results in Semantic Search is a complex task based on various factors like the importance of results, results the user expects, whether it is a comparatively new result etc.

3.3.1 Introduction

In SemRank, the results are ranked on the basis of the predictability of the results the user must be expecting. SemRank offers its users the opportunity to change the effect of the results depending on the depth of the information required. Being able to predict the amount of information the user can gain by making him aware of the existence of the results, gives an estimate of the ability to calculate the amount of information conveyed. This is the underlying principle of SemRank. Assuming ranking schemes, to be suitable or not

can be a drawback, as different applications tend to have different needs.

3.3.2 Working

Keeping in mind the limitations introduced by making assumptions, customizability and flexibility are the two concepts incorporated into SemRank to overcome this limitation. In addition, SemRank provides the user with the functionality of entering a keyword that captures some relevance to the intended results. SemRank assigns the highest rank value to an unpredictable result path in a discovery mode whereas the lowest rank value is assigned in a conventional mode. SemRank values for Semantic Association are computed with the help of annotation of path expression trees and Top-K ordering algorithm. An important point to be noted is that the SemRank ordering is independent of the path length.

3.4 Relation Based Page Rank

The basis of Relation based approach is finding a relation between keywords and concepts of the intended search made by the user, which on the other hand, in traditional search engines are completely based on searching the keywords entered by the user.

3.4.1 Introduction

In Relation Based PageRank it includes only the pages in the list only if the page consists of enough keyword-concept associations linked to the intended user search. The probability of returning page increases, if the number of relations between concepts within the query linked with other concept is larger.

3.4.2 Working

It takes input query as a set of keyboards and finds a logical link between the concepts hidden behind each keyword. The main idea which is finding a relation between keywords and concepts can be carried out in a (semi)automated way or another method can be used i.e. requesting the user to mention the relation and the concept of the keywords which are entered by the user in the search query. The second option which requests the user to enter the relation helps to avoid equivocation and is the main user for its implementation. In this approach the user is asked to enter a keyword and then select a concept from a pull down menu. Consider the following example to understand the concept. Example: The user intends to search and specifies a keyword India and then he or she selects from the pull-down menu any one the concepts like Destination or Country and then second keyword will be Hotel so the concept associated with it will be Accommodation. So the semantic search engine will return pages based on the associations of the keywords and the concepts i.e. it will return pages that has Hotels for accommodations in India consisting on the web pages which are listed. Any other traditional search engine will just return pages without considering the relation between the concepts and the keywords.

3.5 RaRe Rank

In comparison of traditional Information Retrieval methods, link analysis based algorithms have proved to be more effective for ranking documents retrieved from large caches of data like World Wide Web.

3.5.1 Introduction

Rational Research algorithm aka “RareRank”, for Semantic Search is based on the link analysis model. Unlike the traditional IR methods where the focus is on content, in link analysis the focus is on link structures (quality). The RareRank algorithm clearly defines relation between the relevance and the quality score. The underlying principle is that entities such as citations, publishers, authors, journals in combination with the topics in a terminological ontology will be able to simulate an environment suitable for the researchers to conduct research or for any person to be able to carry out search operations.

3.5.2 Working

Unlike the PageRank algorithm, RareRank does not require the value between the documents to be stated explicitly. In addition, navigation between the documents can be indirect by utilizing the established links. The design of RareRank facilitates the promotion of comparatively newly written documents but the ones which are highly relevant to the search query. The working of the algorithm requires a knowledge base and an ontology topic domain. A directed and labeled graph is used to represent a knowledge base in research domain. Then the graph is plugged with relevant domain topic ontology. The graph thus produced is used to simulate the environment as stated. The principle of converge, stated by Markov chain is employed to find the ranking scores. The transition probability is based on the ontology schema and the knowledge base graph. The relationship between the ontological classes and the transition weights is assigned by the ontological classes. The knowledge base graph consists of instances and their relationships generated from the ontology schema.

4. COMPARATIVE STUDY

We have attempted to classify the discussed approaches in such a manner that one can understand each approach and identify their advantages as well as disadvantages. On the basis of a wide array of factors, we have compiled a comparative study and have been displayed in Table 1.

Table 1: Comparative study of various approaches for ranking in Semantic Search

Authors	Approaches	Focus	Association Determination	Architecture	Input	Effectiveness
Rocha et al.[1]	Hybrid Spread Activation	Entity based Ranking	Combination of Clustering measure and Specificity Measure	Stand Alone	Keyword query	Semantically one of the most effective

Anyanwu et al.[2]	SemRank	Relationship based Ranking	Top-K ordering algorithm and Annotation Path Expression	Depends on the architecture of SSARK system	Query and the level of search result required	Effective on small set. Still to be tested on large set
Wei et al.[3]	RaRe Rank	Entity based Ranking	Link Analysis Based	Meta	Keyword query	Very effective when compared to PageRank and HITS algorithm
Lamberti et al.[4]	Relation based Page rank	Relation between keywords & concepts	Page relevance and scoring using query sub graph and ontology graph	Graph Based	Set of keywords, concepts	Effective as it interprets hidden concepts behind keywords
Heflin et al.[5]	SHOE	Relation Based	Navigation of the concept hierarchy	Stand alone	Concepts and property types of ontological structure	Use of semantic mark-up language for annotations to improve efficiency.

5. CONCLUSIONS

Semantics is being incorporated into search engine of major search companies. With the help of this survey paper we aim to elucidate doubts regarding the different approaches for ranking results in Semantic Search. We have classified the five approaches on the basis of parameters that we identified. We have given an overview of each approach to discuss in brief about them as well as try to give a succinct explanation of the working of those approaches. Further the advantages and disadvantages have been stated wherever possible.

The information boom has further aggravated the situation of World Wide Web. Searching has become a complex task. In the purview of overcoming this difficulty has become more important. Semantic Search offers the possible solution to this problem. However, there still exists lot of ground to be covered.

6. REFERENCES

[1] Rocha, Cristiano, Daniel Schwabe, and Marcus Poggi Aragao. "A hybrid approach for searching in the semanticweb." Proceedings of the 13th international conference on World Wide Web. ACM, 2004.

[2] Anyanwu, Kemafor, Angela Maduko, and Amit Sheth. "SemRank: ranking complex relationship search results on the semantic web." Proceedings of the 14th international conference on World Wide Web. ACM, 2005.

[3] Wei, Wang, Payam Barnaghi, and Andrzej Bargiela. "Rational research model for ranking semantic entities." Information Sciences 181.13: 2823-2840, 2011

[4] Lamberti, Fabrizio, Andrea Sanna, and Claudio Demartini. "A relation-based page rank algorithm for semantic web search engines." Knowledge and Data Engineering, IEEE Transactions on 21.1: 123-136, 2009.

[5] J Hefflin, J Handler "Searching the Web with SHOE." Defense Technical Information Center, 2000.

[6] Jindal, Vikas, Seema Bawa, and Shalini Batra. "A review of ranking approaches for semantic search on Web." Information Processing & Management 50.2: 416-425, 2014

[7] Mangold, Christoph. "A survey and classification of semantic search approaches." International Journal of Metadata, Semantics and Ontologies 2.1: 23-34, 2007.

[8] Jin, Yi, Zhuying Lin, and Hongwei Lin. "The research of search engine based on semantic web." Intelligent Information Technology Application Workshops, 2008. IITAW'08. International Symposium on. IEEE, 2008.

[9] Batra, Mridula, and Sachin Sharma. "Comparative Study of Page rank Algorithm With Different Ranking Algorithms Adopted By Search Engine For Website Ranking." International Journal of Computer Technology and Applications 4.1: 8, 2013.

[10] Huiping, Jiang. "Information Retrieval and the semantic web." Educational and Information Technology (ICEIT), 2010 International Conference on. Vol. 3. IEEE, 2010.