Semantic Search Algorithms based on Page Rank and Ontology: A Review

S.G. Choudhary, S.R. Kalmegh, Dr. S. N. Deshmukh

Abstract—As search engines became popular amongst Netizens, a need of semantic search has become a necessity. As the context and semantics of the information in the web pages indexed depends on multiple factor, semantic search has become a complex task. Google Research Lab has worked out many generic algorithms however they are successful in certain conditions only. A focus on personalization of semantic search was given where in one can restrict the domain set and search parameters depending on the personal information. For example, one can build the semantic search techniques based on the domain of some company.

This paper discusses some of the issues handled in semantic search.

Keywords—Ontology, Page Rank, Semantic Search, Web mining.

I. INTRODUCTION

USE of internet has increased in past ten years to the extent that the users are taking help of it for almost all the planning’s in their day to day life. The search engines like Google and Yahoo are so famous that they are in use now and then for searching various type of information available on web. A web has become a largest available data set in public domain to the extent that now-a-days; we are using a term “Information Explosion” as the data indexed by the search engines is so huge.

This explosion of information has brought some side effects of its own. The search using Google is easy but sorting the expected data out of the search results is very difficult. A keyword based search algorithms used in search engines adds more and more confusion in indentifying requisite data. Hence the scientists fear about the hiding of expected information in the large set of relevant and irrelevant information. This becomes worst when the keywords used for searching are unambiguous for ex “lotus” where lotus can be a name of a flower, name of hotel, or some individual etc.

For carrying out the exact search, word sense disambiguates could be used. This process involves the use of other information present in a semantic analysis system. And also it takes help of other words presenting in the sentence and in the rest of the text.

This has proved to be urging of internet user for searching of relevant information. This paper discusses the techniques available for semantic search and also tries to put focus on comparative analysis of them.

II. BACKGROUND AND EXISTING METHODOLOGY

Research related to semantic web search is in process in multiple direction at multiple places the search engines are also devoting there research for semantic search [1], [2]. The normalized Google distance is one of the examples of research direction Google is trying for personalization of search result depending on the ranking techniques and user profile is also considered to be the hot research area in this field.

In this section we have given such available techniques with brief description of it.

A. PSSE (Personalized Semantic Search Engine)

PSSE uses the profiles with the ranking score and ontology for calculation of personalized factor which will help us to get more personalized search result. The architecture of PSSE has two parts, Offline and Online. The Offline part consists of crawling and preprocessing processes. The Online phase includes query processing and result ranking. The fig.1 gives the architecture of PSSE.

The crawler using Offline phase is a multi crawlers which can collect web resources to store in the data set [3], [4]. The preprocessing is divided into three steps. The first step uses indexer and link analyzer to build the graph of the pages that are crawled. The graph so generated is processed in step two by performing link analysis to calculate the reference of pages. The third step is identification of connected components in the graph generated in step one.

In Online phase, GUI is designed for searching the data inserted into it. It is fed to a searcher for searching and retrieval of relevant results. The ranking module is used for ranking retrieval results based on three factors i.e. page reference, relevance of resource content to query terms and Personalization Factor (PF)[5]-[7].
B. Microsearch

Fig. 2 below shows an overview of microsearch system architecture.

The behavior of the system is as follows on the microsearch website [8]: user initially starts same way with Google search engine. The query is given to search engine and the top result are retrieved for display. Along with the regular search results, we also get top results that contain certain types of micro format data. In next step, the metadata is extracted from the displayed result to get the micro format results. After executing Tidy on the pages, the extractor extracts micro formats, linked RDF and RDFa data.

Further, the metadata is collected and stored in a temporary Sesame [9] repository to speed up the further queries. The result is generated and display using the Elmo API to get a Java object model from the RDF data. To get the snippets from metadata the Fresnel API [10] developed by SIMILE is used.

Fresnel transformations are described in declarative manner, giving us other type of properties to display for some classes of objects and some are visualized as link or images. This is known as Fresnel lenses written in RDF using the Fresnel vocabulary. The RDF provides the way to create visualization by inheriting from existing description.

C. Ranking Algorithms

Page Rank is a link analysis algorithm, named after Larry Page and used first by the Google Internet search engine that assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set [Wikipedia].

Ranking algorithms are changing a great deal at the moment. Before, like HITS and Page Rank they looked at how websites and pages were connected and used any information they could gather from those to use as variables. Nowadays we’re seeing research papers come out focusing on how the semantic web, tags, social media can contribute to improved ranking [11].

Some of the Ranking Algorithms are as listed below.

GRank: “exploits the context gained by grouping resources and which improves search for resources”. It also ranks relevant resources according to a keyword query. Like FolkRank it also computes topic-sensitive rankings.

FolkRank: It adapts Personalized PageRank for folksonomies [12] and computes a ranking score of all types of entities in the folksonomy. This algorithm exploits the structure of folksonomy to improve search results. It also computes topic-sensitive rankings. FolkRank algorithm retrieves top a folksonomies element which tends to be in coherent topic area. This leads naturally to the idea of extracting communities of interest from the folksonomies, which are identified by their top tags and the most influential persons and resources. Making such communities explicit will make interested users to find it and participate [13].

GFolkRank and GFolkRank+: Group-sensitive FolkRank (GFolkRank) is an adaptation of the 2 above. It “interprets groups as artificial, unique tags”. Graph-based ranking algorithms, which extend FolkRank and turn it into a group-sensitive algorithm in order. In addition to the above, GFolkRank+ is capable of propagating tags, which have been assigned to a group, to its resources.

SocialPageRank: The SocialPageRank algorithm [14] is motivated by the observation that there is a strong interdependency between the popularity of users, tags, and resources within a folksonomy. For example, resources become popular when they are annotated by many users with popular tags, while tags, on the other hand, become popular when many users attach them to popular resources [14].

SocialSimRank: SocialSimRank and FolkRank both base on the PageRank algorithm. Regarding the underlying random surfer model of PageRank [15], a remarkable difference between the algorithms relies on the types of links that can be followed by the “random surfer”. SocialPageRank restricts the “random surfer” to paths in the form of resource-user-tag-resource tag-user, whereas FolkRank is more flexible and allows paths like resource-tag-resource.
It ranks resources and tags respectively. It computes static, global rankings independent of the context. It determines the similarity between tags or query terms based on the tag assignments.

III. ONTOLOGY BASED IR MODEL

Propose model for exploitation of ontology based KBs to search over large knowledge base in the order of GBs or TBs consisting of thousands ontology instances. The approach of retrieval is based on adaption of classic vector space model [16], annotation weighting algorithm and ranking algorithm. In classic vector model, keywords are assigned weight to indicate that some words are better in documents than other.

In weighing annotations, annotations are assigned a weight that shows how relevant the instance is considered to be for the document meaning. Weights are computed automatically by an adoption of TF-IDF algorithm [16], based on the frequency of occurrences of the instances in each document.

Ontology-based approaches allow for sophisticated semantic search but impose query syntax more difficult to handle [17] presented an approach for translating keyword queries to DL conjunctive queries using background knowledge available in ontologies. It also presents an implementation which shows that this interpretation of keywords can then be used for both exploration of asserted knowledge and for a semantics-based declarative query answering process. Castells [18] proposed a model for the exploitation of ontology-based knowledge bases to improve search over large document repositories. The retrieval model is based on an adaptation of the classic vector-space model, including an annotation weighting algorithm, and a ranking algorithm. Semantic search is combined with conventional keyword-based retrieval to achieve tolerance to knowledge base incompleteness. Experiments are shown where our approach is tested on corpora of significant scale, showing clear improvements with respect to keyword-based search.

Julian Seidenberg, Alan Rector [19] evaluated several algorithms for extracting relevant segments out of large description logic ontologies for the purposes of increasing tractability for both humans and computers. The segments are not mere fragments, but stand alone as ontologies in their own right. This technique takes advantage of the detailed semantics captured within an OWL ontology to produce highly relevant segments.

After the list of documents is formed, the search engine computes a semantic value between the query and each document. For example, the query of user is for special offers for summer holidays in Kerala, a document that shows one such offers will get similarity value in the order of 1/n, where n is total number of registered offers in knowledge base that match the query. Only a document that displays nearly all offers could get similarity 1. If the knowledge in KB is incomplete the semantic ranking algorithms performs vary poorly, query will return less results than expected, and the relevant documents will not be retrieved, or it will get a much lower similarity value than expected. To overcome this, ranking model combines the semantic similarity measure with similarity measure of a keyword based algorithm.

IV. ANALYSIS

Personalized Semantic Search gives us the results which are up to the mark but within some particular domains. The ontology used is PSSE depends on the specific domain of internet. A generic approach may require combinations of data sets and related ontologies which may make the system very complex.

Microsearch algorithm uses Google search engine to get the hits and then process it for semantic analysis with the help of repositories and rules. All this processing is at a cost of more clock cycles. Also, as Google top results are based on several factor, the data Google is referring, will always not hit the requisite data.

The algorithms based on PageRank have revealed different results on different data. When tested on tagged data GFolkRank performs well than FolkRank and SocialPageRank as GFolkRank algorithm ranks first tag consistently.

However, for untagged data, FolkRank is proved to be better than SocialSimRank algorithm and achieved better precision when tested for Rank 3.

Ontology based IR models uses OWL. A comparative study shows that ontology models have generated more relevant data as compared to others. But efficient models are still in awaited.

V. CONCLUSION

Semantic Search is necessity of the word as it is very difficult to get the relevant information from the information ocean. Several algorithms are designed in this domain and researchers are still designing new one. However, an efficient, generic and fast algorithms are still is in want.

REFERENCES


[17] Ontology-based Interpretation of Keywords for Semantic Search by Thanh Tran, Philipp Cimiano, Sebastian Rudolph, Rudi Studer.

[18] An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval Pablo CASTELLS, Miriam FERNÁNDEZ, and David VALLET.

[19] Web Ontology Segmentation: Analysis, Classification and Use Julian Seidenberg, Alan Rector Medical Informatics Group, School of Computer Science University of Manchester, Manchester, United Kingdom.