International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463
Vol. 4 Issue 3, March-2015, pp: (269-273), Impact Factor: 1.252, Available online at: www.erpublications.com

Semantic Web Technologies and Algorithm

Gaurav[1], Mr. Lokesh[2]

[1]M.Tech. Student, RIEM, Rohtak, Haryana
[2]HOD, CSE Dept. RIEM, Rohtak, Haryana

Abstract: The World Wide Web (WWW) allows the people to share the information from the large database repositories globally. The amount of information grows billions of databases. There are various search engines available today, but it is very difficult to retrieve the meaningful information. However, semantic web technologies are playing a major role to overcome this problem in search engines to retrieve meaningful information intelligently. This thesis presents survey on the search engine generations and the role of search engines in intelligent web and semantic search technologies. In our literature survey on semantic web search engines, most of the search engines search for keywords to answer the queries from users. The search engines usually search web pages for the required information. However, they filter the pages from searching unnecessary pages by using advanced algorithms.

General Terms: Semantic web, Web mining and semantic web approaches.

Keywords: Semantic Web Mining, Web 3.0.

---

1. INTRODUCTION

The Semantic Web aims to build a common framework that allows data to be shared and reused across applications, enterprises, and community boundaries. It proposes to use RDF as a flexible data model and use ontology to represent data semantics. Currently, relational models and XML tree models are widely used to represent structured and semi-structured data. But they offer limited means to capture the semantics of data. An XML Schema defines a syntax-valid XML document and has no formal semantics, and an ER model can capture data semantics well but it is hard for end-users to use them when the ER model is transformed into a physical database model on which user queries are evaluated. RDFS and OWL ontologies can effectively capture data semantics and enable semantic query and matching, as well as efficient data integration. The following example illustrates the unique value of semantic web technologies for data management.

Concept Based Search

A lot of semantic web search works are opted to add semantic annotations to data in order to increase the search precision and recall on that data. Consequently, the exploitation of the semantics contained in the concepts, their relationships and instances. The data in semantic web can be divided into two categories: ontological (concept) and instance data. The data which have an interest to user are instances relative to ontological class.

Relation Based Search

In this type of search method, relations between query terms are inferred from knowledge bases to aid the retrieval process. In Kim and Ontolook, the entity relations are processed using word combination. Ontolook replaces query terms by concept pairs and sends these pairs to the knowledge base to extract all relations which have been asserted in the knowledge base. The main idea of this method is the identifications of relations among concepts included in the keywords of user query.

Ranking Based Search

This approach starts with an initial set of relations (properties) by adding hidden relations, which can be inferred from the query. Then the inferred relations will be computed as the ratio between relation instances linking concepts specified in the

user query and the relation instances in the semantic knowledge base. In addition, all the relations of interest are requested to be specified by user.

Mining Based Search

Exploiating semantics for web mining or mining the semantic web based on background knowledge (ontologies or other) can be used to improve the process and results of web mining. The quality of semantic search depends largely on the quality and the coverage of knowledge base. The knowledge concerned by a mining process refers to the hidden knowledge neither asserted in knowledge base nor derived using logical inference with rules. Such knowledge can only be derived from large amount of data by using some methods of analysis techniques.

2. Improved Semantic Web Algorithm

Semantic web documents are the mixture of classes and relationships among them. They hold the metadata that describe the digital objects identified by URI. Conceptually, they hold two kinds of information: schema and instances. Fully understanding this document requires complex set of tasks depending on the granularity of retrieved information by search engines. The information in these documents is represented as the graph structure. The relevant information is organized as graphs on the basis of developed schema. The schema is developed using ontology.

In this work, these algorithms (Concept matching algorithm, retrieving inbound anchors algorithm, and minimal answer algorithm) are applied one by one following some pattern so as to get benefit of all these algorithms, as all these algorithms are designed for semantic web search. But these algorithms individually have some different benefits like minimal answers algorithm gives the precise and in short hand answer to user query, while retrieving inbound anchors algorithm gives more similar links to user query. So, in this work these semantic web algorithms are applied to show that proposed algorithm gives better results than keyword based algorithm as well as Yandex (a Semantic search engine). In this work, the concept of ontology is used, which is a lexical database for the english language. Wordnet, DBPedia, YAGO, text runner, SUMO are all ontologies, in which various entities and facts about those entities are already stored.

Steps in the proposed algorithm are:

Step 1: Input: Natural Language Query.

Step 2: Translate the query into its linguistic triple form using Linguistic Component.

Step 3: Maps the terms of each linguistic triple to semantically relevant ontology Entity using Mapping Component.

Step 4: After that apply module to eliminate redundant links in the retrieved Ontologies [15].

Step 5: Select the ontological triples that best represent the user's query using Triple Similarity Service Component.

Step 6: Output: list of semantic entities retrieved from different ontologies and KBs

Firstly, the linguistic component analyzes the natural language query and translates it into its linguistic triple form. e.g a query "What are the cities of Spain?" has the linguistic triple (<what-is, cities, Spain>). In the second step, the ontology discovery sub module identifies the set of ontologies likely to provide the information requested by the user. To do so, it searches for approximate syntactic matches within the ontology indexes, using not just the linguistic triple terms, but also lexically related words obtained from wordnet and from the ontologies, used as background knowledge sources. e.g the term cure match with the concepts cure, heal, treat etc. Once the set of possible syntactic mappings have been identified, the semantic filtering sub module checks its validity using a wordnet-based filtering methodology. This methodology is based on a semantic similarity measure between the set of synsets t obtained for the query term T and the set of synsets c obtained for the matched concept C. To do so, the measure considers the path distance (depth) and the shared

Semantic Similarity (t, c) = t ~c = (2 x depth (C.P.I (t, c))) / (depth (t, c) + 2 x depth (C.P.I (t, c)))

To elicit the sense of a mapped concept C with respect to a query term T, we intersect (1) SC,T, the set of synsets of C semantically similar to T, with (2) SHC, the set of synsets of C that are semantically similar to any synset of its ontology ancestors. Obviously, if this intersection is empty it means that the sense of the concept in the ontology (2) is different from the sense defined by the query term T (1), and therefore that mapping should be discarded.

$$SC,T = \{c \in SC \mid \exists t \in ST \text{ such that } t \sim c\} \quad (1)$$

$$SHC = \{c \in SC \mid \forall R ((R>C) \rightarrow (\exists r \in SR (c \sim r)))\} \quad (2)$$

After this process, a set of entity mapping tables is generated where each table links a query term with a set of concepts mapped in the different domain ontologies. After this process, the triple similarity service module takes as input the previously retrieved entity mapping tables and the initial linguistic triples and extract, by analyzing the ontology relationships, a small set of ontologies that jointly covers the user query. The output of this module is a set of triple mapping tables where each table relates a linguistic triple with all the equivalent ontological triples. Using these triples the information of the knowledge bases is analysed to generate the final answer.

3. Results with Improved Semantic Web Algorithm

The results with the improved algorithm are derived using the three types of queries:

1. Single Keyword Queries
2. Multiple Keyword Queries
3. Complex Queries

Results for Single Keyword Queries

The following queries has been taken as examples of single keyword queries in all the search engines:

Q1.1 Database
Q1.2 Multimedia
Q1.3 Software
Q1.4 Program
Q1.5 Hardware

A. Response Time

Table 1: Response Time (in ms)

	Response Time in milliseconds		
Search Queries	**Google**	**Semantic Search Engine (Yandex)**	**Proposed System**
Q1.1	0.19	0.15	0.13
Q1.2	0.17	0.14	0.13
Q1.3	0.18	0.17	0.17
Q1.4	0.20	0.14	0.11
Q1.5	0.20	0.15	0.12
Total	0.94	0.75	0.66
Average	0.188	0.15	0.132

In table 1, single keyword queries are taken and response time (in ms) is shown for Google, Yandex and proposed system. It has been observed that proposed system takes minimum time for the query Q1.4 which is 0.11ms, followed by the

Yandex's time which is 0.14ms for the query Q1.2 while google takes minimum time for the query 1.2, which is 0.17ms. In brief, in all the queries(from Q1.1 to Q1.5), proposed system takes less time to response the user query.

B. Estimated Result Count

Table 2: Estimated Result Count

	Estimated Result Count		
Search Queries	**Google**	**Yandex**	**Proposed System**
Q1.1	1, 440, 000, 0000	265, 000, 000	99, 600, 000
Q1.2	1, 150, 000, 000	189, 000, 000	107, 000, 000
Q1.3	4, 670, 000, 000	474, 000, 000	3, 56, 000, 000
Q1.4	622, 000, 000	452, 000, 000	286, 000, 000
Q1.5	273, 000, 000	228, 000, 000	125, 000, 000
Total	21,115,000,000	1,608,000,000	8,74,000,000
Average	4,223,000,000	3,21,600,000	1,74,800,000

In table 2, it has been observed that Google shows very large estimated count, which is 1, 440, 000, 0000 for the query Q1.1, while for Yandex, it is 474,000,000 for the query Q1.3, followed by the proposed system, which is 356,000,000 for the query Q1.3.

Graphs

The following graphs are shown for the given scenarios:

2.2.2.1 Scenario I: Graphs for Single Keyword Queries

A. Response Time

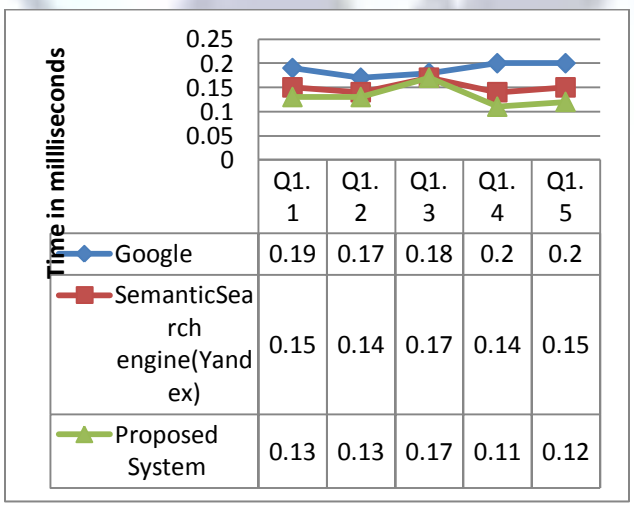

Figure 1: Response Time (in ms)

In figure1, on X-axis, single keyword queries are taken and on Y-axis, response time(in ms) is shown for Google, Yandex and proposed system. It has been observed that proposed system takes minimum time for the query Q1.4 which is 0.11ms, followed by the Yandex's time which is 0.14ms for the query Q1.2 while google takes minimum time for the query 1.2, which is 0.17ms.

B. Estimated Result Count

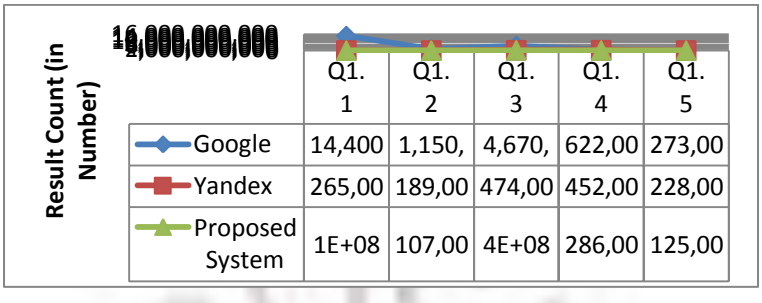

Figure 2 : Estimated Result Count

CONCLUSION

The information available on the web is unstructured, disorganized, dynamic and heterogeneous in nature. Moreover the process of retrieval is highly affected by the ill formed queries put up by the average user. Today's search engines returns too many results which are not necessarily relevant to the user's need. Usually, a user has to traverse several search result pages to get to the desired results. Many commercial search results such as Google (based on Page Rank) and Yahoo are being used by people across the globe, but the relevancy of documents returned in the search engines results still lacks. That's why rigorous researches are being carried out in the field of information retrieval. The objective of a search engine must be to satisfy user's information need in a lucid way. Semantic search techniques are undeniably useful in this context. They have potential to produce more relevant and precise results according to user query as shown in this thesis work.

REFERENCES

[1]. Agualimpia, W. R., LopezPellicer, F. J., MuroMedrano, P. R. and Nogueras Iso, J., 2010. "Exploring the Advances in Semantic Search Engine", International Symposium on Distributed computing and artificial Intelligence, Springer, Vol. 79, pp. 613-620.
[2]. Beel, J. and Gipp, B., 2009. "Google Scholar's Ranking Algorithm: The Impact of Citation Counts (An Empirical Study)", Proceedings of the 3rd IEEE International Conference on Research Challenges in Information Science, Vol. 3, pp. 439–446.
[3]. Berkan, R.C. and Pulatkonak, M., 2004. "Yadex Search Engine", International Journal on Advances in Information Sciences and Service Sciences, Vol. 2, No. 01, pp. 102-112.
[4]. Brin, S. and Page, L., 1998. "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Seventh International World Wide Web Conference, Vol. 2, pp. 107-117.
[5]. Cole, R. and Hariharan, R., 1998. "Approximate string matching: a simpler faster algorithm", Proceedings of ACM- SIAM SODA, Vol. 2, pp. 463-472.
[6]. Ding, L., Finin, T., Joshi, A., Pan, R., 2004. "Swoogle: A search and metadata engine for the semantic web", Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management, Vol. 3, pp. 652-659.
[7]. Finin, T., Mayfield, J., Fink, C. and Joshi, A., 2005. "Information retrieval and the semantic web", Proceedings of the 38th International Conference on System Sciences, Vol. 2, pp. 113-118.
[8]. Ge, J. and Qiu, Y., 2009. "Concept Similarity Matching Based on Semantic Distance", Fourth International Conference on Semantics, Knowledge and Grid, Vol. 1, pp. 380-383.