Journal of Computer Science 10 (9): 1776-1781, 2014
ISSN: 1549-3636

doi:10.3844/jcssp.2014.1776.1781 Published Online 10 (9) 2014 (http://www.thescipub.com/jcs.toc)

ONTOPARK: ONTOLOGY BASED PAGE RANKING FRAMEWORK USING RESOURCE DESCRIPTION FRAMEWORK

[1]Yasodha, S. and [2]S.S. Dhenakaran

[1]Department of Computer Science,
Government Arts College for Women, Pudukkottai, Tamilnadu, India
[2]Department of Computer Science and Engineering, Karaikudi, Tamilnadu, India

Received 2014-03-31; Revised 2014-04-01; Accepted 2014-04-24

ABSTRACT

Traditional search engines like Google and Yahoo fail to rank the relevant information for users' query. This is because such search engines rely on keywords for searching and they fail to consider the semantics of the query. More sophisticated methods that do provide the relevant information for the query is the need of the time. The Semantic Web that stores metadata as ontology could be used to solve this problem. The major drawback of the PageRank algorithm of Google is that ranking is based not only on the page ranks produced but also on the number of hits to the Web page. This paved way for illegitimate means of boosting page ranks. As a result, Web pages whose page rank is zero are also ranked in top-order. This drawback of PageRank algorithm motivated us to contribute to the Web community to provide semantic search results. So we propose ONTOPARK, an ontology based framework for ranking Web pages. The proposed framework combines the Vector Space Model of Information Retrieval with Ontology. The framework constructs semantically annotated Resource Description Framework (RDF) files which form the RDF knowledgebase for each query. The proposed framework has been evaluated by two measures, precision and recall. The proposed framework improves the precision of both single-word and multi-word queries which infer that replacing Web database by semantic knowledgebase will definitely improve the quality of search. The surfing time of the surfers will also be minimized.

Keywords: Semantic Web, Ontology, Resource Description Framework, Knowledgebase

1. INTRODUCTION

The Web contains heterogeneous information such as text, hyperlinks and multimedia. For information retrieval from the Web users rely on traditional search engines that do not provide any means of considering the semantics of data. So, handling keywords with multiple semantics is often an omitted task of search engines. For example, the keyword Principal would mean Head of the institution in one context and Amount invested in another context. This disparity could not be dealt with by search engines and they provide information related to both contexts when the term Principal is given as search keyword.

Another problem with search engines is Web spamming. Due to Web spamming, irrelevant Web pages are boosted to top-order and relevant Web pages do not receive due importance.

To solve these problems, Semantic Web has emerged that helps to provide the most relevant results for the users' query. The Semantic Web is an extension of the current Web in which the semantic annotation of each page is stored along with the contents of the Web page (Davies *et al*., 2003). The semantics of the different terms in a particular domain are provided as ontology. So ontology based frameworks could be designed that possess knowledge about the user query, annotated Web pages and the underlying ontology.

Corresponding Author: Yasodha, S., Department of Computer Science, Government Arts College for Women, Pudukkottai, Tamilnadu, India

Four types of technologies are available for building the Semantic Web: Metadata, Ontology, Logic and Agents (Antoniou and Harmelen, 2004). In this study an ontology based framework for ranking Web pages has been proposed, implemented and tested. This framework was implemented in JAVA and ontology construction was done using Resource Description Framework (RDF). The performance of the framework was evaluated using two metrics, precision and recall.

1.1. Ontology

The term ontology denotes a formal and explicit specification of a shared conceptualization (Borst, 1997). Ontology includes terms and their relationships. The term denotes important concepts of the domain. For example, in a university domain, students, courses, faculty members and disciplines are some of the concepts. The relationships denote hierarchies of classes. Ontologies are helpful for the navigation and organization of Websites. They are also helpful for increasing the precision of Web searches.

Ontology is a knowledge representation method. It uses classes and properties for organizing the knowledge and represents the data or image in a structured way (Magesh and Thangaraj, 2013). The ontology makes it possible to search both explicit and tacit knowledge, thereby bridging the gap between the explicit and tacit knowledge. The advantages of ontology are knowledge sharing, logic inference and reuse of knowledge (Vadivu and Hoper, 2012). Two types of ontologies exist: (i) General-purpose ontologies and (ii) Domain-specific ontologies. General-purpose ontologies aim to provide conceptualizations of general notions. Domain-specific ontologies are intended for sharing concepts and relations in a particular area of interest (Al-Safadi and Al-Abdullatif, 2010).

There are four important components of ontology. They are:

- Concepts-A concept denotes a set or class of entities or `things' within a domain. For example:

  Vice-Chancellor is a concept within the domain of University

- Relations-Relations indicate the interactions between concepts or a concept's properties. For example

  Vice- Chancellors areappointedby the Governor

- Instances-Instances are the `things' indicated by a concept. For example

  Malala is an instance of the concept student

- Axioms-Axioms are used to constrain values for classes or instances For example

  Students securing less than 50% of marks should reappear

1.2. Resource Description Framework (RDF)

RDF is a World Wide Web Consortium (W3C) specifications originally designed as a metadata data model. RDF is a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web. It stores metadata about files and other machine-accessible resources (Gauthami Latha *et al*., 2011). RDF documents consist of three types of entities:

- Resources-Resources may be Web pages, parts or collections of Web pages, or any real-world objects that are not directly part of the WWW. In RDF, resources are always addressed by URIs
- Properties-Properties are specific attributes, characteristics, or relations describing resources
- Statements-Each statement consists of (Resource, Property, Value) triples. In the RDF graph example shown in **Fig. 1**
  Ponting is a resource
  <plays> is a property
  The string « Cricket » is a value

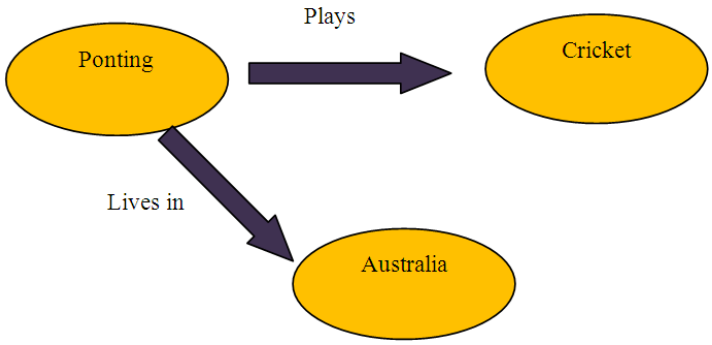

Fig. 1. RDF graph example

RDF data has become a reliable source of information for many applications. For example, in resource discovery to provide better search engine capabilities. RDF with digital signatures is the key in building the "Web of Trust" for electronic commerce, collaboration and other applications.

2. MATERIALS AND METHODS

We proposed a new framework named ONTOPARK for ranking relevant Web pages. ONTOPARK was designed using RDF ontologies. The proposed framework was designed as an extension of the traditional Vector Space Model of information retrieval. It was combined with ontology, the Semantic Web technology that enables meaningful information retrieval from the Web. The framework works in three phases: Preprocessing, Ontology Construction and Ranking. The framework design is shown in **Fig. 2**.

2.1. Phase I-Preprocessing

In this phase, the framework accepts the query from the user and extracts Web links from Web database using Google. The top 30 Web links ranked by Google are taken for preprocessing. Then it preprocesses the query as well as the snippets and contents of each Web page by applying preprocessing steps like removal of insignificant words like a, an, the, by, with and removal of suffix. For example, the words talk, talking and talkative are reduced to their root word talk by suffix removal.

2.2. Phase II-Ontology Construction

After preprocessing the query, snippets and the contents, RDF knowledgebase is constructed for each query. RDF files are created for the top 30 Web links whose page rank of Google is non-zero. The RDF files are created by combining the Web Link (URL), title, preprocessed snippet and the preprocessed contents corresponding to each Web link. The collection of these RDF files forms the RDF knowledgebase for that query. This RDF knowledge base is used in the next phase for ranking.

2.3. Phase III-Ranking

Ranking is based on the adaptation of the Vector Space Model of information retrieval. In the Vector Space Model, term weights are computed for query terms by counting the number of occurrences of the term in the documents of the Web database. But in the proposed framework, term weights are computed for query terms that appear in the RDF files of the RDF knowledgebase. Term weight is computed by an adaptation of the TF-IDF algorithm, where TF denotes the Term frequency and IDF denotes the inverse document frequency. Using this term weight, relevance score is computed to measure the similarity of the query to each RDF file in the RDF knowledgebase. Ranking is done based on this relevance score:

$$\text{Precision} = \frac{\text{Total relevant for each query}}{\text{Total retrieved for that query}}$$

$$\text{Recall} = \frac{\text{Total retrieved for each query}}{\text{Total available for that query}}$$

Mean Precision/Recall = Average Precision/Recall of Single-Word and Multi-Word queries.

Mean Average Precision/Recall = Average of Mean Precision/Recall of Single-Word and Multi-Word queries.

Consider Knowledgebase K with RDF files r_1, r_2,…r_m. The framework accepts a query Q = $\{x_1…x_n\}$ containing the terms $\{x_1…x_n\}$. The answer to the query is a list of the top n documents. The term frequency tf (x,r) is the number of times that the term x appears in RDF file r. The document frequency df (x,K) is the number of RDF files in K that contain x.

The weight W (x,r) of a term x in an RDF file r is computed as:

$$W(x,r) = tf(x,r)\ X idf(x,K)$$

where, tf(x,r) is the normalized frequency of term x in RDF file r which is computed as:

$$tf(x,r) = \frac{freq(x,r)}{\max\{freq(y,r)\}}$$

where, freq(x,r) is the number of occurrences of the term x in r.

max {freq(x,r) } is the frequency of the most repeated term y in RDF file r.

The inverse document frequency idf(x,K) is computed as:

$$idf(x,K) = \log\frac{N}{df(x,K)}$$

where, N is the set of all RDF files in the knowledgebase and df(x, K) is the number of RDF files in Knowledgebase K annotated with x.

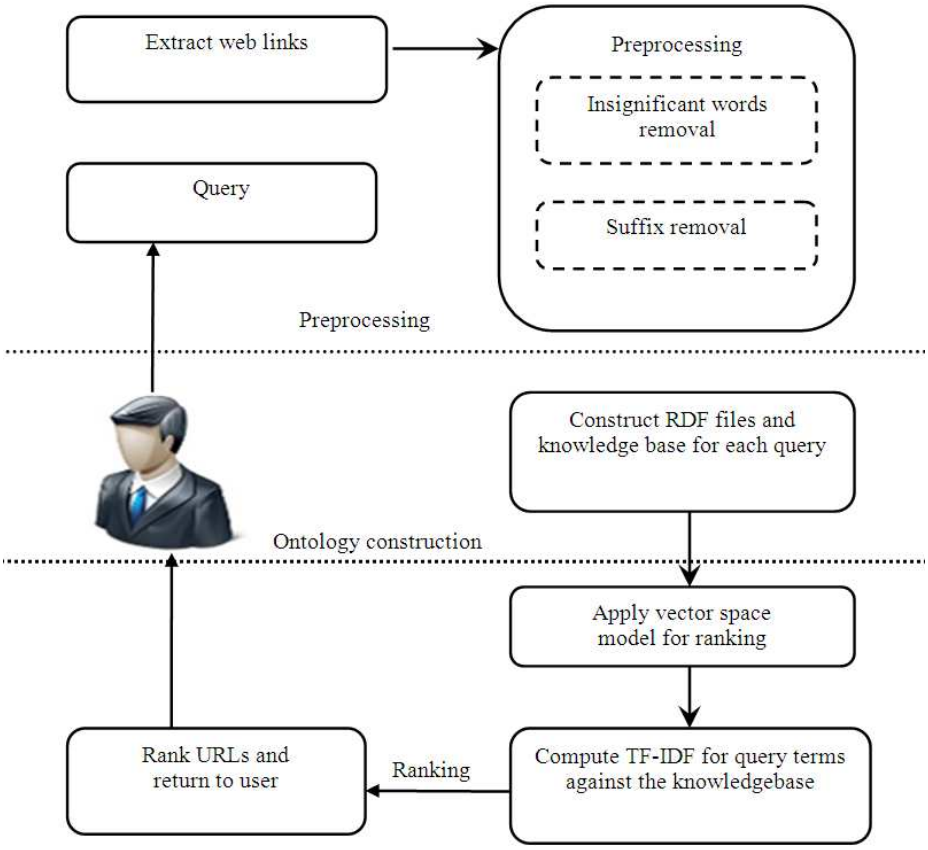

Fig. 2. Architecture of proposed framework

The documents are ranked according to a relevance score Score(Q, r),which is the relevance of an RDF file r to the query Q:

$$\text{Score}(Q,r) = \sum_{X \in Q,r} W(x,r).\text{In}\frac{|K|+1}{df(x,K)}$$

where, |K| = m is the size of the Knowledgebase K.

2.4. Evaluation Measures

The page ranks of Web pages produced by any search engine or framework could be evaluated by two measures: Precision and recall. Precision is the measure of accuracy. It measures the relevance of Web pages with respect to the total retrieved. Recall measures the quantity of Web pages retrieved with respect to the total available.

3. RESULTS

The framework was implemented in JAVA and the screenshots were designed using Net Beans IDE. Ontology engineering was done using RDF. The framework was tested with single word and multi word queries. The performance was evaluated by two metrics precision and recall. The results were compared to that of Google. The results are tabulated in **Table 1-3**. The page ranks produced by ONTOPARK and Google for the keyword "Data mining" is given in **Table 4**.

4. DISCUSSION

The proposed framework produces better precision values though for a few queries, the recall values of Google are better. This is because only the Web pages for which the Google rank is non-zero are considered for RDF file construction and ranking. One can find the page ranks of Google by installing Google's tool bar or by page rank check tools like www.prchecker.info. When compared to the ranking of Google, ONTOPARK produces better ranking because in the PageRank algorithm of Google, the number of hits to Web pages are also considered for ranking.

Table 1. Mean average precision and recall

	Proposed	Google
Mean precision of single word query	0.76	0.72
Mean precision of multi word query	0.80	0.70
Mean average precision	0.78	0.71
Mean recall of single word query	0.62	0.64
Mean recall of multi word query	0.71	0.52
Mean average recall	0.67	0.58

Table 2. Precision and recall (Single-word Query)

	Proposed		Google	
Query	Precision	Recall	Precision	Recall
Networking	0.6	0.83	0.5	0.80
Data	0.9	0.77	0.8	0.63
Java	0.8	0.62	0.6	0.82
Laptop	0.7	0.71	0.8	0.50
Apple	1.0	0.50	0.9	1.00
Canon	0.9	0.67	0.8	0.38
Satellite	0.7	0.71	0.8	1.00
Resort	0.9	0.44	0.8	0.38
Inverter	0.7	0.43	0.6	0.33
System	0.8	0.50	0.9	0.56

Table 3. Precision and Recall (Multi-word Query)

	Proposed		Google	
Query	Precision	Recall	Precision	Recall
Data Mining	0.9	0.80	0.8	0.50
Colleges for doing MBA	0.8	0.90	0.6	0.50
How far is Tagore University	0.9	0.44	0.8	0.62
Research scope in India	0.5	0.85	0.6	0.67
Star hotels in Chennai	0.9	0.22	0.7	0.50
Flights to Malaysia	0.8	0.88	0.7	0.50
Symptoms of dengue	1.0	0.60	0.8	0.50
How is dollar value determined	0.8	0.75	0.5	0.20
What is the use of PAN card	0.8	0.75	0.3	0.67
Online shopping in Chennai	0.7	0.86	0.5	0.50

Table 4. Page Ranking for the query ‘Data mining

Query	URLs	Proposed Rank	Google Rank
Data Mining	http://datamining.typepad.com/	1	6
	http://en. Wikipedia.org/wiki/Data_mining	2	6
	http://www.eco.utexas.edu/~norman/BUS.FOR/course.mat/Alex/	3	3
	http://www.oracle.com/technetwork/database/options/odm/index.html	-	0
	http://www.kdnuggets.com/publications/	4	5
	http://www.webopedia.com/TERM/D/data_mining.html	5	4
	http://www.kmining.com/	6	4
	http://www.autonlab.org/tutorials/	7	5
	http://www.investopedia.com/terms/d/datamining.asp	8	3

This paved way for Web spamming, the illegitimate means of boosting page ranks. For example, as we could see in **Table 4**, the page rank of Google for the Web link http://www.oracle.com/technetwork/database/options/odm/index.html corresponding to the keyword “Data mining” is 0, but this Web link has been ranked in top order. As irrelevant ranking of Web pages are prevented, the precious surfing time of the surfers will be definitely reduced.

5. CONCLUSION

We designed a framework that constructs RDF knowledgebase for each query. The RDF files in the knowledgebase were annotated with semantic information which helped for the meaningful retrieval of information. The limitation with this framework is that RDF files were created only for the top 30 Web pages. The number of RDF files created for each query should be increased so as to include more number of relevant Web pages for ranking. Though the area of Semantic Web has got high focus now-a-days, there is still there is a long way to go in the area of Semantic Web and research in this particular area should also be encouraged. In future more sophisticated ontology languages such as OWL may be used for ontology engineering to exploit the maximum benefits of using such languages.

6. ACKNOWLEDGEMENT

The researcher S.Yasodha would like to acknowledge University Grants Commission (UGC) of India, for their financial support extended to this project under the Minor Research Project Scheme (LINK F. 3923/11 UGC-SERO).

7. REFERENCES

Antoniou, G. and F.V. Harmelen, 2004. A Semantic Web Primer. 1st Edn., The MIT Press, London, England. ISBN-10: 0-262-01210-3, pp: 8-15.

Borst, W., 1997. Construction of Engineering Ontologies for Knowledge Sharing and Reuse. PhD Thesis, University of Twente.

Davies, J., D. Fensel and F.V. Harmelen, 2003. Towards the Semantic Web Ontology-Driven Knowledge Management. 1st Edn., John Wiley and Sons, England. ISBN-10: 0470 84867 7, pp: 4.

Gauthami Latha, A., R.N. Raju and C. Satyanarayana and Y. Srinivas, 2011. Hotel management system: An E-commerce application using resource description framework RDF. Int. J. Comput. Sci. Inform. Technol., 2: 2267-2272.

Magesh and Thangaraj, 2013. Comparing the performance of semantic image retrieval using SPARQL query, decision tree algorithm and lire. J. Comput. Sci., 9: 1041-1050. DOI: 10.3844/jcssp.2013.1041.1050.

Vadivu, G. and S.W. Hoper, 2012. Ontology mapping of Indian medicinal plants with standardized medical terms. J. Comput. Sci., 8(9) : 1576 – 1584. DOI: 10.3844/jcssp.2012.1576.1584.

Al-Safadi, L.A.E. and N.A.O. Al-Abdullatif, 2010. Educational Advertising Ontology: A domain-dependent ontology for semantic advertising networks. J. Comput. Sci., 6: 1070-1077. DOI: 10.3844/jcssp.2010.1070.107.