PERFORMANCE EVALUATION OF A RELATION BASED PAGE RANKING ALGORITHM USED FOR IMPROVING SEARCH RESULTS

Sumera Hayat 1  MS CS,  Nadeem Qazi 2  Associate Professor

1 2 Faculty of Engineering, Sciences & Technology. IQRA University, Karachi, Pakistan.

Abstract—Search engines help us mine the web and get what we are looking for. But the problem with today’s search engines is that, it gives us most reputable results but not really the most relevant ones. The search criteria of these traditional web search engines is based on keyword matching or to some higher extent (like in Google), matching words similar or in the same context to the query word and then ranking the matched pages with PageRank Algorithm, based purely on link analysis. PageRank despite being quiet efficient in sorting the results to be presented to the user, still results in a number of irrelevant pages. To overcome this problem and provide relevant results, the Relation Based Page Ranking Algorithm has been introduced recently. In this paper, we evaluate the performance of the Relation Based Page Ranking Algorithm on some academic web pages annotated with an Ontology specifically built for the web page’s domain. Ranking of web pages through this algorithm is based upon a probability measure that checks the relevancy of a page to the query. Finally, we compare the results of our experiments with that of Google’s results generated for the same query.

Keywords: Semantic web, PageRank, Ontology, Relevance Ranking, Annotations.

1. INTRODUCTION

Search Engines are developed to help the users search for what they are looking for in the web’s data repository. Search Engines work with the help of predefined automated software programs that are known as Spiders or bots and are used for crawling, that is, going through the entire website (all the pages) and records down the content in the form of databases known as indexes. These search engine indexes are invoked to display the webpage content, when the user defines a query in the search engine interface and the query is parsed and checked against the search engine database to get the best results possible [1]. The matching criteria of the existing search engines is usually based on keyword matching, that is, matching the query keywords with the web page’s keywords stored in the indexes of the search engines. There is no architecture on the web 2.0 [14] to store the relations a user has in mind while typing the query [2]. At the time of query processing, the relations are lost and the given keywords are treated as individual keywords, hence creating the major problem of isolated keyword matching. Even the ranking of the retrieved web pages has no account for relations and is purely based on link analysis like in Google’s PageRank [3][4] and some on page relevance factors [5]. There should be some way to measure these relations and to make keywords search more meaningful that is, incorporating Semantics into the search engines working. Though Google has injected some of the very important semantic web technologies [6] such as latent semantic indexing [15] and displaying rich snippets [16], but it still processes and displays some irrelevant results on the very first page.

In this paper, we prove that the Relation Based Page Ranking Algorithm proposed in [8] is a solution to the above mentioned inconsistencies of the existing search engines. The Ranking strategy is based on the ontology data, user query and the page annotations, which will be exploited to measure the relevance of a web page to a given query. The Ranking criterion proposed specifically for the Semantic Search Engines, in no ways eliminates the use of the existing ranking algorithms like PageRank, which checks for the repute of pages. In fact, this Relation Based Ranking Algorithm can be used in conjunction with the PageRank algorithm to give most reputable and most relevant results on a semantic search engine.

2. RELATED WORKS IN RANKING ALGORITHMS FOR THE SEMANTIC WEB

The aim of this paper is to make use of the relations embedded as annotations within a web page and the query concepts for creating a ranking strategy capable of assigning a ranking score better than the ones done in today’s search engine’s ranking algorithms. This idea of making use of the ontology based semantic meta data for ranking web pages is not new [9][10][11]. How ever, these previous approaches did not consider the semantic relations which are said to be the key component of the Semantic Web. To make optimum use of the Semantic Web content marked up, there should be approaches available that takes into account the relations and the semantic web associations between content in annotated web pages that can be used for ranking and retrieving data [2][12][13].

The most relevant work with respect to this paper would be [8] and [2]. The basic idea of [2] is that if a graph based web page annotation can be provided, where concepts and relations are modeled as vertices and weighted edges, respectively, it becomes possible to define a series of cuts removing less relevant concepts from the graph. This allows for the generation of a so
called candidate relation –keyword set (CRKS) to be submitted to the annotated database, which can reduce the presence of uninteresting pages in the result set. However, the effectiveness of this approach is strongly limited as there is no kind of ranking strategy involved. In [8] we see an extension to this approach which relies on the assumption that for providing effective ranking, the search engine logic should only need to know the structure of the underlying ontology and the web page to be ranked in order to compute the relevance score.

3. OVERVIEW OF THE RANKING STRATEGY

For the Semantic web to work optimally and give the results actually anticipated by the users, we need to not only add in the so called semantic web plug ins, but to change the entire infrastructure of the current web search engines. The Relation Based Page Ranking Algorithm for the Semantic Search Engine is proposed to solve the issues related to the ranking techniques of today’s web 2.0, like PageRank used by Google. To prove that this Algorithm works better than the existing search engines ranking systems we have created a data set to reproduce the ranking results generated by Relation Based Page Ranking Algorithm in [8]. We first present the real time environment in which this algorithm will be applicable in future.

3.1 Infrastructure of a Semantic Search Engine

The Relation Based Page Ranking Algorithm cannot be used along with the current web search engines infrastructure as the current search engines lack to store the relations between query keywords and the annotated web documents which are the basis of Relation Based Page Ranking Algorithm. In order to understand the ranking algorithm itself, we first need to understand the environment where it will actually function in real time scenarios. The prototype of the Relation Based Page Ranking Algorithm presented in [8] is shown in figure 3.1. In this paper, we have not focused on how to create this tiny Semantic web environment, but to evaluate the performance of the Ranking Logic of this Semantic Web, hence limiting our focus to the Ranking Module of the figure below.

Figure 3.1 Prototype Architecture for the Relation Based Page Ranking Algorithm

3.1 Scenario for the Evaluation of the Algorithm

When searched through Google (on 12 January 2012, at 3:44 pm) about Dr Alam Raza’s Research work who is working at IQRA University with the following keywords, “Dr Alam Raza” “Research” “IQRA University”, the search results displayed the web page (Faculty Web page of IQRA university) on top of the list and gives the profile page on 7th position in the list, which actually had the research details of Dr Alam Raza. The third page which is also less relevant than the profile page but is much closer to the query than the faculty web page is ranked 3rd and the least relevant page that contains nothing about Dr Alam Raza’s Research work is ranked 2nd by Google. The keywords were merely matched on the basis of keywords density and latent semantic indexing, with no importance given to the relations in between the keywords which a user had in mind. For instance in this case, the relations in our mind were Dr Alam Raza has Research some Research, who is working at IQRA University. The ranking of these web pages was based of the link analysis PageRank Algorithm which gave no importance to the relations. The faculty web page having a PageRank 4/10 [18], was ranked first as the other pages had PageRank less than this.

3.2 Creating OWL ontology

Ontology is a way of representing things of the world, called entities, into a concept based method which defines a kind of taxonomy or hierarchy. It gives a common vocabulary for the people and machines to share knowledge of a particular domain [19]. So we created an ontology using Protégé 4.1[20][21] for manually annotating some web pages of IQRA University. The ontology shown by a graphical option (Ontoviz) in Protégé 4.1 can be seen in figure 3.2.

Figure 3.2 IQRA University Ontology Onto viz view in Protégé 4.1

The above graphical representation of the ontology can be written in an ontology language [22] understood by machines, which will be the language for annotating the web pages as well. We have extracted the RDF/XML code of this ontology written in a web ontology language [23] from Protégé 4.1 after its creation and verification by a reasoner. A portion of the RDF/XML code of the IQRA University ontology can be seen in figure 3.
3.3 Manual Annotation of web pages
The task performed to make web page data available and readable to the machines is known as Annotations or Semantic Markup. The term annotate simply means to attach data to some other piece of data [24]. Semantic annotations can be given to a web document in many ways. The traditional way is with the help of the tools like OntoMat annotizer[25], Annotea [26], SMORE [27] etc. Other ways are through Semantic Wikis, Semantic blogs, tagging with the help of RDFa, microformats and embedding RDF [7] meta data with the help of ontology vocabulary. With the help of our IQRA University Ontology we manually annotated 4 web pages of IQRA University and the RDF/XML code generated for one of these web pages using Protégé Ontology RDF/XML code is shown in figure 3.4.

3.4 Graph Based Notation and Methodology
Starting from the ontology defined for a domain, a graph based representation can be designed where OWL classes are mapped into graph vertices and OWL relation properties are mapped into graph edges. A link between concepts in the graph shows a relation between the concepts and has some weight on the basis of the given relations. Like wise, we formulate the query and annotated web pages into their respective graphs using the graph based notations given below in table 3.1. Given an ontology graph G and a query sub graph GQ, it is possible to define a ranking strategy capable of assigning each page including queried concepts a relevance score based on the semantic relations available among concepts within the page itself (thus neglecting the contribution of the remaining Web pages). The proposed ranking strategy assumes that given a query Q, for each page p, it is possible to build a page sub graph GQ,p using a methodology that is similar to the one used for G and GQ and exploiting the information available in page annotation A. All these graphs will help us implement the Relation Based Page Ranking Algorithm.

3.5 Graph Based Formulization
So now using the graph theory we formulate the ontology graph for the IQRA University ontology created in section 3.2. We start with the ontology graph to be built over the part of IQRA University Ontology shown below in figure 3.5 (a)(b):
Assuming a Semantic Search Engine, we pass a query containing three keywords and three concepts associated with it. Keywords: “Dr Alam Raza” “Research” “IQRA University” and the associated Concepts: “Teaching Staff” “Research” “Universities”. So now we formulate the query sub graph based on the query and web page 1’s sub graph based on the query sub graph in figure 3.5 (d) and (e).

### 3.6 Page Relevance Score and Ranking

For each page sub graph, all the possible spanning forests [28][29] (both constrained and unconstrained) are generated with either progressively removing edges of all the spanning forests to get constrained forests or by adding new edges to all the constrained forest of length 1 until a spanning forest is obtained [30]. We use the first approach for our experiments. Total number of spanning forests can be identified with the Cayley’s formula $[31] = n^{n-2}$. Considering page 1’s sub graph in figure 3.5 (e), we find out the total spanning trees of this sub graph, $n^{n-2} = 3^{3-2} = 3$

So we have total 3 spanning trees for page 1’s sub graph with edges $n-1 = 2$ edges.

With the help of the original page spanning forest we can generate all the constrained page spanning forests. The constrained page spanning forests has edges less than the original spanning forest and if equal than it will be a spanning forest not a constrained forest. So all the possible constrained page spanning forests with edges $= 1 = 1$ are in figure 3.7.

**Figure 3.6 All possible spanning trees of Page 1 Sub graph**

**Figure 3.7 All possible constrained page spanning forests**

We find the probability $P(Q, p, 1)$ of a web page with constrained page spanning forest of length 1 in the following way:

Constrained Page Relevance Score of Page 1

$$P(Q, p, 1) = \sum_{i=1}^{n} P(r_{ij}, p_i) P(SF^{(i)}_{p_i}, p(1))$$

$$P(Q, p, 1) = P(r_{12}, p_1) P(SF^{(1)}_{p_1}, p(1)) + P(r_{13}, p_1) P(SF^{(2)}_{p_1}, p(1)) = 0.5555$$

Similarly, we compute the relevance score for the other web pages as well. The results of the ranking are presented in the table below with each page showing its ranks of Google and the
Relation Based Page Ranking Algorithm, along with the constrained page relevance score.

<table>
<thead>
<tr>
<th>Google Ranking</th>
<th>Our Ranking</th>
<th>Web page details</th>
<th>Relevance Score</th>
</tr>
</thead>
</table>
| 1              | 3           | Faculty | Iqra University  
www.iqra.edu.pk/?page_id=1  
210  
... University Ranking. Click here for details Business  
School's Research Ranking of Pakistan ... Dr. Alam Raza.  
PhD, (Eco.). MSc (Eco.) ... Iqra University, Karachi ... | 0.5555         |
| 2              | 4           | Program Teams | Iqra University  
www.iqra.edu.pk/?page_id=2  
070  
Click here for details Business  
School's Research Ranking of Pakistan ... Dr. Alam Raza; Mr.  
Muhammad Ahsan ullah Khan Durran; Mr. Faisal K Qureshi | 0.1111         |
| 3              | 2           | IURC Organogram | Iqra University  
www.iqra.edu.pk/?page_id=1  
667  
Latest University Ranking. Click here for details Business  
School's Research Ranking of Pakistan ... DR. ALAM RAZA,  
Social Sciences. DR. ASADULLAH LARIK ... | 0.6111         |
| 7              | 1           | IU Learning Management System  
iulms.edu.pk/profile/publicprofile.php?UserID=353-08-1005  
Picture of Dr. Alam Raza ... At present, working at Iqra  
University, Defence View  
Campus (Business Administration Department),  
Karachi, Pakistan, as an ... | 0.6666         |

Table 3.2 Accuracy of the Ranking Algorithm checked over 4 web pages

4. Conclusion

It can be clearly seen that the profile page which had the research details about Dr Alam Raza has been ranked 1st by the Relation Based Page Ranking Algorithm applied. Hence it is proved that using the relations between concepts embedded in a web page (as semantic annotations) a ranking criterion formulated known as the Relation Based Page Ranking Algorithm, results in much more accurate results than Google and matched user needs to a greater extent.

To benefit from the accuracy of Relation Based Page Ranking Algorithm, we need to opt for the Semantic Web Environment, which needs to be built on top of the existing web and not by just adding the semantic web technologies into the existing web. The search engines today are becoming more and more semantic and are trying to provide users with accurate results. But in order to implement this novel Ranking Strategy, the existing search engines would require this module to be implemented in a Web 3.0 environment, as it’s the best platform for this ranking strategy to work to its full potential.

REFERENCES

http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html
[8] A Relation-Based Page Rank Algorithm for Semantic Web Search Engines. Fabrizio Lamberti, Member, IEEE, Andrea Sanna, and Claudio Demartini, Member, IEEE
[10] Swoogle: A Semantic Web Search and Metadata Engine. Li Ding, Tim Finin, Anupam Joshi, Yun Peng, R. Scott Cost, Joel Sachs, Rong Pan, Pavan Reddivari, Vishal Doshi
[16] Introducing Rich Snippets,  
http://googlewebmastercentral.blogspot.com/2009/05/introducing-rich-snippets.html
[17] H. Knublauch, Protégé, Stanford Medical Informatics,  
http://protege.cim3.net/file/pub/ontologies/travel/travel.owl
http://www.prchecker.info/check_page_rank.php
[22] Choosing an Ontology Language. Anna V. Zhdanova, Uwe Keller

[23] Web Ontology Language: OWL. Grigoris Antoniou₁ and Frank van Harmelen₂ Department of Computer Science, University of Crete, ga@csd.uoc.gr₁ Department of AI, Vrije Universiteit Amsterdam₂.

[24] What are Semantic Annotations?? Eyal Oren₁, Knud Hinnerk Möller₁, Simon Scerri₁, Siegfried Handschuh₁, and Michael Sintek₂.₁ Digital Enterprise Research Institute, National University of Ireland, Galway {eyal.oren, knud.moeller, simon.scerri, siegfried.handschuh}@deri.org.₂ German Research Center for Artificial Intelligence (DFKI) sintek@dfki.uni-kl.de

[25] Onto Mat Annotizer


[27] SMORE.
http://annotation.semanticweb.org/Members/lago/AnnotationTool.2003-08-25.4401

[28] Spanning trees and Spanning forests
www.cafed.sssup.it/~giulio/software/spanntree/spanning_tree.html

[29] Spanning Trees and Optimization Problems. Bang Ye Wu, Department of Computer Science and Information Engineering, Shu-Te University, Yen-Chau, Kaohsiung County, Taiwan 824. Kun-Mao Chao, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan 106.

[30] RELATION BASED SEMANTIC WEB SEARCH ENGINE S. Raja Ranganathan₁, Prof. M. Sadish Sendil₂, Dr. S. Karthik₃