International Journal of Information Processing, 7(2), 20-31, 2013
ISSN : 0973-8215
IK International Publishing House Pvt. Ltd., New Delhi, India

Mathematical Model of Semantic Look - An Efficient Context Driven Search Engine

Leena Giri G[a], Srikanth P L[a], S H Manjula [a], K R Venugopal [a], L M Patnaik[b]

[a]Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore 560 001 India, Contact: leenagiri@gmail.com.

[b]Honorary Professor, Indian Institute of Science., Bangalore.

The World Wide Web (WWW) is a huge conservatory of web pages. Search Engines are key applications that fetch web pages for the user query. In the current generation web architecture, search engines treat keywords provided by the user as isolated keywords without considering the context of the user query. This results in a lot of unrelated pages or links being displayed to the user. Semantic Web is based on the current web with a revised framework to display a more precise result set as response to a user query. The current web pages need to be annotated by finding relevant meta data to be added to each of them, so that they become useful to Semantic Web search engines. Semantic Look explores the context of user query by processing the Semantic information recorded in the web pages. It is compared with an existing algorithm called OntoLook and it is shown that Semantic Look is a better optimized search engine by being more than twice as fast as OntoLook.

Keywords : Ontology, RDF, Semantic Web.

1. INTRODUCTION

Semantic Web (Web 3.0) is the proliferation of unstructured documents of the web to a "web of data" [1]. In traditional web architecture there is less emphasis on meta data of the web document during the data collection phase of the search engine and the concentration is more on classic approaches like Information Retrieval and Natural Language Processing. It is difficult to know the context or the role played by the web document designed for such approaches [2][3][4]. This is overcome by Semantic Web where enhanced version of meta data are embedded in the web pages as RDF [5] and Ontology [6]. Ontology defines the concepts and the relations between these concepts. RDF (Resource Description Framework) describes the web document in the form of triplets. Every RDF triplet is a composition of *subject*, *predicate* and *object*. *Subject* is an entity to be described, *object* is an entity which describes the subject and *predicate* is a relationship between *subject* and *object*; essentially every *predicate* describes the different context of the web page playing multiple roles. Both Ontologies and RDF are embedded in web pages forming the semantic annotation of a web page.

1.1. Motivation

The existing search engines interpret the keywords of a user query in isolation without considering wholly, the context of the search query. Because of this, most of the results retrieved is irrelevant to the user query. This hits the performance and accuracy of search engines. The main purpose of providing multiple keywords is to make search based on a particular context. It is to say that nothing exists without context or relation. As an example, consider a scenario where a user has submitted the keywords "Ashoka+Bangalore+Hotel" with the intention to search for Hotel Ashoka in Bangalore. Traditional web search engines return all the web pages containing the keywords "Ashoka, Bangalore and Hotel" without considering the context of the user query. Most of the web pages are irrelevant to the user query; where some pages may provide information on

for a user query is established by extracting the relations among the supplied keyword. This is performed by *Semantic Look*.

The entire application is developed on *LAMPP* environment with*PHP* as underlying language for business logic. As shown in *Fig 3* the Semantic Look and Ontolook is compared with respect to the number of relations to be processed for different sets of keywords and concepts provided by the user.The difference in the number of sub graphs processed by OntoLook and Semantic Look is given in Table 7.
Since in every sub graph high ranked edges are retained and only the selected less ranked edges are pruned, the number of sub graphs to be processed is less in Semantic Look compared to Ontolook. As shown in *Table 7*, the number of relations to be processed in Semantic Look is less than half of the number of relations processed by Ontolook as depicted in *Figure 3*. Every sub graph produces large number

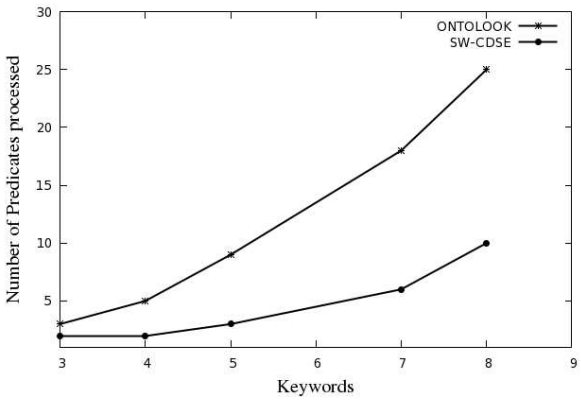

Figure 3. Keywords Predicates Processed

of duplicate RDF triplets which is submitted to the Ontobase to fetch URLs for every sub graph and intersection of these URL sets produce the distinct set of URLs as a result set for the user. The search time here includes the time for pruning the selected less ranked edges from the Ontology graph, producing the RDF triplets and database communication time for fetching the URLs set from it. From *Table 7* and *Figure 3* it is shown that number of sub graphs produced in Semantic Look is less compared to Ontolook and therefore the number of RDF triplets produced in Semantic Look is less which in turn reduces the search time as compared with Ontolook. *Table 8* shows the number of RDF triplets processed and search time taken by Ontolook and Semantic Look.

8. CONCLUSIONS

Search engines in the current web architecture will not consider the semantics role played by web pages in different context. The new generation of web *i.e.*, Semantic Web (web 3.0) considers this context information by recording the semantic information in the form of Ontologies and RDFs. A proof of concept called Semantic Look is proposed to produce relevant web pages by filtering unnecessary web documents from the result set.

Semantic Look extracts the semantics of the user query to know the context of user search. This work is based on the prototype called OntoLook which performs the exhaustive search of all the sub graphs of Ontology graph to produce URL set. Semantic Look is an optimized search engine compared to OntoLook which prunes less weighted edges from the OntoLook to produce less number of sub graphs for processing.

Even though the number of sub graphs processed by Semantic Look is less as compared with OntoLook the number of RDF triplets produced will be huge and therefore in future work Semantic Look should be designed to run on the clusters of nodes using Map-Reduce Framework. Further optimization is achieved by running the crawler and pruning logic on the cluster. Since semantic information is embedded in the web page by the author and it is assumed to be true there is a chance of misleading the search engine by embedding false semantic information.

REFERENCES

1. W3 Consortium, Semantic Web, *http://en.wikipedia.org/wiki/Semantic Web.*
2. T Priebe, C Schliiger and G Pernul. A Search

Table 6
Sub Graphs Processed for a Particular Combination Keywords and Relations

No.of Keywords		No.of Relations		No.of Sgraphs processed	
OLook	SLook	OLook	SLook	OLook	SLook
8	8	25	10	5200300	252
7	7	18	6	48620	20
5	5	9	3	26	3
4	4	5	2	10	2
3	3	3	2	3	2

Table 7
No. of RDF Triplets Produced and Search Time to process them for Combination Keywords and Relations

Sub graphs processed		RDF triplets produced		Process Time	
OLook	SLook	OLook	SLook	OLook	SLook
5200300	252	701345778	81144	710039	34.4076
48620	20	5209920	4320	21.0912	1.874
126	3	6832	354	3.2049	0.216094
10	2	244	120	0.12911	0.0634
3	2	48	46	0.02599	.02399

Engine for RDF Metadata, *Proceedings of the* 15^{th} *International Workshop Database and Expert Systems Applications*, pages 168-172, 2004.

3. T Berners-Lee, J Hendler and O Lassila. The Semantic Web, *Scientific American*, 284(5):34-43, 2001.
4. A Gomez-Perez and O Corcho. Ontology Languages for the Semantic Web, *IEEE Intelligent Systems*, 17(1):54-60, January-February 2002.
5. D Beckett. RDF/XML Syntax Specification (Revised), *http://www.w3.org/TR/2004 /REC-rdf- syntax-grammar-20040210/, 1994.*
6. John Hebler, Matthew Fisher, Ryan Blac and Andrew Perez-Lopez. Semantic-Web Programming, *Wiley India Pvt.ltd*, Third Edition, 2009.
7. Yufei Li, Yuan Wang and Xiaotao Huang. A Relation-Based Search Engine in Semantic Web, *IEEE Transactions on Knowledge and Data Engineering*, 19(2):273-281, February 2007.
8. J Cho, H Gareia-Molina, and L Page. Efficient Crawling through URL Ordering, *Proceedings of the* 8^{th} *International World Wide Web Conference*, May 1999.
9. Li Ding, Finin, Joshi, Pan, Scott Cost, Sachs, Doshi, Reddivari, and Y. Peng. Swoogle, a Search and Metadata Engine for the Semantic Web, *Proceedings,* 13^{th} *ACM Conference on Information and Knowledge Management (CIKM 2004)*, Nov. 2004.
10. Noy, Sintek, Decker, Crubezy, Fergerson, and M A Musen. Creating Semantic Web Contents with Protege-2000, *IEEE Intelligent Systems*, 16(2):60-71, March-April 2001.
11. Lastra L M J and Delamer M. Semantic web Services in Factory Automation: Fundamental Insights and Research Roadmap, *IEEE Intelligent Systems*, 6(6):1-11, February 2006.
12. Yi, Jin, Zhuying Lin and Hongwei Lin. The Research of Search Engine Based on Semantic Web, *IEEE Intelligent Systems*, 360-363, December 2008
13. Alexander Maedche, Boris Motik, Ljiljana Stojanovic, Rudi Studer, and Raphael Volz. Ontologies for Enterprise Knowledge Management, *IEEE Computer society*, 2003.
14. Wang Young-gui and Jia Zhen. Research on Semantic Web Mining, *IEEE conference Publications*, 1:67-70, June 2010.
15. Ramesh Singh, Dhruv Dingra and Aman

Arora. SCHISM - A Web search engine using semantic taxonomy, *IEEE POTENTIALS*, September-October 2010.

16. Mohammad Farhan Husain, James McGlothlin, Mohammad Mehedy Masud, Latifur R. Khan and Bhavani Thuraisingham. Heuristics-Based Query Processing for Large RDF Graphs using Cloud Computing, *IEEE Transactions on Knowledge and Data Engineering*, 23(9), September 2011.

Leena Giri G is currently an Associate Professor in the Department of Computer Science, Dr. Ambedkar Institute of Technology, Bangalore. She obtained her Bachelor of Engineering from SJCE, Mysore. She received her M.Tech Degree in Computer Science and Engineering from IIT Mumbai. Her research interest is in the area of Semantic Web.

Srikanth P L received his Master's degree from the Department Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. His research interest is in the area of Web Technology, Semantic Web and Cloud Computing.

S H Manjula is currently the Chairman, Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. She obtained her Bachelor of Engineering and Masters Degree in Computer Science and Engineering from University Visvesvaraya College of Engineering. She was awarded Ph.D in Computer Science from Dr. MGR University, Chennai. Her research interests are in the field of Wireless Sensor Networks and Data mining.

K R Venugopal is currently the Principal, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. He obtained his Bachelor of Engineering from University Visvesvaraya College of Engineering. He received his Masters degree in Computer Science and Automation from Indian Institute of Science Bangalore. He was awarded Ph.D in Economics from Bangalore University and Ph.D in Computer Science from Indian Institute of Technology, Madras. He has a distinguished academic career and has degrees in Electronics, Economics, Law, Business Finance, Public Relations, Communications, Industrial Relations, Computer Science and Journalism. He has authored 39 books on Computer Science and Economics, which include Petrodollar and the World Economy, C Aptitude, Mastering C, Microprocessor Programming, Mastering C++ and Digital Circuits and Systems *etc.*. During his three decades of service at UVCE he has over 350 research papers to his credit. His research interests include Computer Networks, Wireless Sensor Networks, Parallel and Distributed Systems, Digital Signal Processing and Data Mining.

L M Patnaik is currently Honorary Professor, Indian Institute of Science, Bangalore, India. He was a Vice Chancellor, Defense Institute of Advanced Technology, Pune, India and was a Professor since 1986 with the Department of Computer Science and Automation, Indian Institute of Science, Bangalore. During the past 35 years of his service at the Institute he has over 500 research publications in refereed International Journals and Conference Proceedings. He is a Fellow of all the four leading Science and Engineering Academies in India; Fellow of the IEEE and the Academy of Science for the Developing World. He has received twenty national and international awards; notable among them is the IEEE Technical Achievement Award for his significant contributions to High Performance Computing and Soft Computing. His areas of research interest have been Parallel and Distributed Computing, Mobile Computing, CAD for VLSI circuits, Soft Computing and Computational Neuroscience.