Proceedings of
INTERNATIONAL CONFERENCE ON COMPUTING, COMMUNICATION AND ENERGY SYSTEMS
(ICCCES-16)
In Association with IET, UK & Sponsored by TEQIP-II
29th -30th, Jan. 2016

Paper ID: CSEIT09

FEATURE EXTRACTION TECHNIQUES USING SEMANTIC BASED CRAWLER FOR SEARCH ENGINE

Poonam P. Doshi,
Research Scholar and Associate Prof., JSCOE
Dr. Emmanual M
HOD IT, PICT Pune

Abstract--The vast amount of data on the World Wide Web has resulted in convergence powerful analytical technologies, namely the Semantic Web. The Semantic Web has allowed for evolution of linked multidimensional data that allows for querying information on the web using their semantics or meaning and not just a list of key words. The paper discussed how Semantic Web technologies evolved from the traditional Extract, Transform and Load (ETL) based to the more automatic mapping of multidimensional data. Primary objective of the proposed research is to improve effective and accurate information discovery over the internet, to study the vocabulary of the mining service, to enable the crawler to work for an uncontrolled web and improve the efficiency of the algorithm.

Index Terms— Crawlers, Semantic web based search engines, Ontology,

I. INTRODUCTION

Internet has turned into an imperative wellspring of data nowadays. Due to its dynamic nature, proficient pursuit component is imperative. Web search tools are utilized to concentrate data from web and crawlers are main piece of web indexes. We can easily get the information by crawling the surface web using traditional web agents.

There are many types of crawlers available as below –

1. Focused web crawler
2. Incremental crawler
3. Distributed crawler
4. Parallel Crawler

A focused crawler is a web crawler that gathers Web pages which has some specific property, by carefully prioritizing the crawl frontier and handling the hyperlink exploration process. Few predicates can be based on simple, deterministic and surface properties. For example, a crawler's intension may be to traverse pages from only the .jp domain. A focused crawler must calculate the probability that an unvisited page will be relevant before actually downloading the page. The anchor text of links was the possible predictor. And this approach was taken by Pinkerton in crawler developed in the start of Web world. The type of focused crawlers which is semantic focused crawler, for the selection and categorization purpose, domain ontologies are used to represent topical maps and link Web pages with relevant ontological concepts. Ontologies may be adaptively updated in the crawling process. An ontology-learning-based crawler was introduced by Dong [7] et al using support vector machine to renovate the content of ontological concepts when crawling Web Pages.

The main advantage of using ontologies is the formalized semantics. Semantic web based search engines employ ontologies in a particular domain to enhance the performance of information retrieval process. The ability to deduce additional facts based on the axiomatic content of ontology can be important from a research point of view. A reasoned can automatically infer new statements without writing specific code.

II. RELATED WORK

Semantic focused crawler has the capacity to cross the web and download related data. Semantic technologies have been broadly used in the field of industrial automation. A study by Kamal Taha [1], the methods for analysis for molecular research has become a de facto standard. Gene set enrichment method have been developed for analyzing. Gene Ontology (GO) terms expounding the set S, which may lead to erroneous results On the other hand, most current similarity measure methodologies may return an enhanced gene set, where some of the genes in the set are annotated with the functions of GO terms connected by GO relations that do not represent existence dependency between the terms. With respect to RG Finder, it embraces the idea of existence dependency for determining the functional and semantic relationships of GO terms/genes. Anuar [3] the existing trademark search systems are primarily based on text-based retrieval. Such systems search for trademarks that match some or all words in a string text query. There is need of extending the current approach to include retrieving trademarks with phonetic similarities and integrating their previous work on visual similarity with their new algorithms for conceptual and phonetic similarity.

Self-adaptive semantic focused crawler – SASF crawler Hai Dong [5], with the purpose of precisely and efficiently discovering, formatting and indexing mining service information over the internet This framework incorporates the technologies of semantic focused crawling and ontology learning, in order to maintain the performance of this crawler, regardless of the variety in the web environment. A study by Dong [7] found that most of the crawlers during this domain build use of ontology have to represent the information underlying topics and web documents. Crawling performance crucially depends on the quality of ontology's web the reason behind this concept its weights are heuristically predefined before being applied to calculate the relevancy scores of web page. In learnable focused crawling approach, this issue is considered. Artificial Neural Network is used to solve problems which cannot be defined in series of steps like recognizing patterns, classification into groups, series prediction and data mining. So focused crawling can be applied here.

Domain specific ontology can be used to construct ANN and used for classification of web pages. Quality of ontology is based on two issues. The first issue is that, as it is well known that ontology is the formal representation of specific domain knowledge [10] and ontology's are designed by domain experts. There may be differences between understanding of the domain knowledge and the domain knowledge that exists in the real world. The second issue is that knowledge is dynamic and is constantly changing, compared with relatively static ontology's.

Researchers have started to pay attention to enhancing semantic focused crawling technologies by integrating them with ontology learning technologies to solve problems in ontology. Also they are trying to enhance the performance of semantic focused crawlers. Wong says, the goal of ontology learning is to semi automatically extract facts or patterns from a corpus of data and turn these into machine readable ontology's [11]. There are various techniques designed for ontology learning. Few of them are statistics based techniques, natural language processing based techniques, logic based techniques, etc. These techniques are classified into supervised techniques, semi-supervised techniques, and unsupervised techniques from the perspective of learning control.

Sr No	Title	Algorithm	Concept	Advantage	Limitation
1.	Scheduling algorithms for web crawling	Crawling the large sites first	Crawling starts with the sites with the large number of pending pages, i.e. webpages for crawling.	Large web site crawled first	When important pages exist in short web site, then this is crawled latter.
2.	Effective Page Refresh Policies for Web Crawlers ACM Transactions on Database Systems.	Breadth first search algorithm.	Starts at the root URL and searches the all then neighbors URL at the same level.	Well suited for situations where the objective is found on the shallower parts in a deeper tree.	It will not perform so well when the branches are so many in a game tree.
3.	Artificial Intelligence illuminated.	Depth first search algorithm.	Starts at the root URL and traverse depth through the child URL.	Well suited for such problems.	When the branches are large then this algorithm takes might end up in an infinite loop.

Ontology learning based techniques can be used to solve the issue of semantic focused crawling, by learning new knowledge from crawled documents and integrating the new knowledge with ontology's in order to constantly refine the ontology's.

III. SEMANTIC WEB OVERVIEW

The semantic web achieves linking of the data using Vocabularies, also referred to as Ontologies. The primary role of Ontologies is to help with integration of data, remove ambiguities and to organize knowledge. "Ontology is defined as a formal, explicit specification of a shared conceptualization". The architecture of the Semantic Web [1] is illustrated in Fig 1.

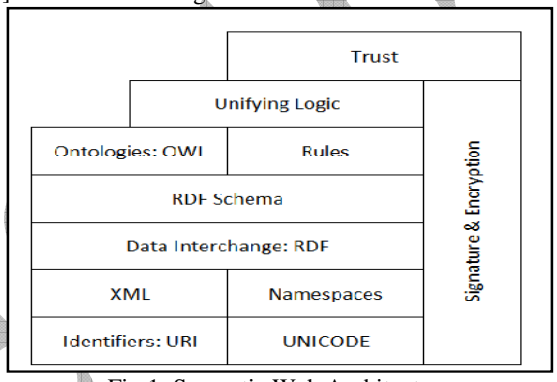

Fig 1: Semantic Web Architecture

The extensible markup language allows for proper syntax to model document contents. RDF [12] gives a graph based model to describe objects and their relationships. RDF Schema [13] provides support for a vocabulary and axioms for describing properties and classes of the RDF-based resources. OWL [14] adds additional vocabulary support for describing properties and classes in order to construct ontology.

Compared to traditional web technologies which focus mainly on representing data, the Semantic Web provides for a more machine-readable platform that allows for the extraction of information about web resources and relationships between various heterogeneous resources.
The primary role of the Semantic Web technologies is to define a common vocabulary of standard and constraints (inferring rules) in order to create a semantic metadata.

The semantic web data should follow four principles [6]:

- Use of Uniform Resource Identifiers (URIs) to identify object.
- Use of Hypertext Transfer Protocol (HTTP) to facilitate searching for objects
- Us of se the Resource Description Framework (RDF) [3] format as a standard to provide descriptive information about an object

To link URIs to others in order to link data on the web

OBJECTIVES OF THE PROPOSED STUDY

Primary object of the proposed research is to improve effective and accurate information discovery over the internet, to study the vocabulary of the mining service, to enable the crawler to work for an uncontrolled web and improve the efficiency of the algorithm.

IV. RESEARCH METHODOLOGY

Development of lead search engine uses the approach like depth first search, crawling, filtering, ranking and AI algorithms. Proposed semantic web crawler is based on supervised and unsupervised ontology-learning, which focuses on the implementation of online Page importance calculation algorithm. In this algorithm, the crawler will download web pages with higher cashes in each stage and cash will be distributed between the pages it points when a page is downloaded. The main idea of this crawler is to construct an artificial neural network (ANN) model to determine the relatedness between a web links and ontology. Semantic focused crawling technology is used to solve the issues of heterogeneity, ubiquity and ambiguity of mining service information, and ontology learning technology is used to maintain the high performance of crawling in the uncontrolled web environment. Intelligent web agent for B2B portal has three layer architecture as follows

- User Interface: Web Pages - This is the first point of contact for the end user of B2B portal. Seed URLs - Seed URLs are the input URLs from end user which user wants to crawls.

- Web Layer: Crawling - Actual crawling of those URLs given by end users will be happening in this module. Term Processing - This module will process the terms which are searched by the user. Term Extraction - From term processing, important terms are extracted here. Filtering- After getting result from result processor, Integrator will filter the results as per user need. Indexing - Results will be indexed accordingly. Auto Formatting - Different formats will be used to display results.

- Database Layer: String Matching - In Ontology learning, String matching will be done and if new string is found important then that will be sent to updater. Updater - Updater will be updated database repository for new strings found. Data Repository - This is the actual database in which all data will be stored. Result Processing - Result processing module will take result from database repository for processing and sending it further for integration.

WEB CARWLER STRATEGIES

1. Breadth First Search Algorithm

This algorithm aims in the uniform search across the neighbor nodes. It starts at the root node and searches the all the neighbor nodes at the same level.

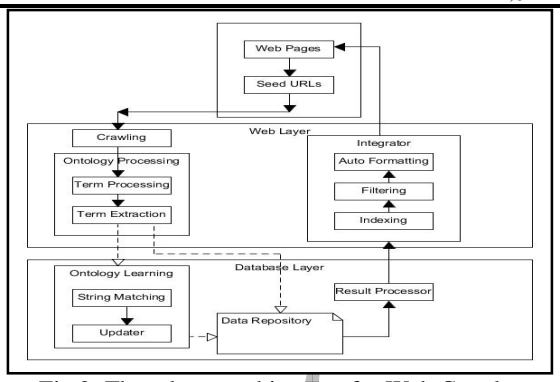

Fig 2: Three layer architecture for Web Crawler

If the objective is reached, then it is reported as success and the search is terminated. If it is not, it proceeds down to the next level sweeping the computation time. BEMADS and GEMADS these two algorithms are used based on Gaussian mixture model. Both resumes data into sub cluster and after that generate Gaussian mixture. These two algorithms run several orders of magnitude faster than maximum with little loss of search across the neighbor nodes at that level and so on until the objective is reached. When all the nodes are searched, but the objective is not met then it is reported as failure. Breadth first is well suited for situations where the objective is found on the shallower parts in a deeper tree. It will not perform so well when the branches are so many in a game tree especially like chess game and also when all the path leads to the same objective with the same length of the path. A distributed BFS for numerous branches using Poisson random graphs and achieved high scalability through a set of clever memory and communication optimizations

1. Put all the given seeds into the queue;
2. Prepare to keep a list of "visited" nodes (initially empty);
3. As long as the queue is not empty:
   a. Remove the first node from the queue;
   b. Append that node to the list of "visited" nodes
   c. For each edge starting at that node:

i. If the node at the end of the edge already appears on the list of "visited" nodes or it is already in the queue, then do nothing more with that edge;

ii. Otherwise, append the node at the end of the edge to the end of the queue.

2. Depth First Search Algorithm

This powerful technique is systematically traversing through the search by starting at the root node and traverse deeper through the child node. If there are more than one child, then priority is given to the left most child and traverse deep until no more child is available. It is backtracked to the next unvisited node and then continues in a similar manner. This algorithm makes sure that all the edges are visited once breadth [15]. It is well suited for search problems, but when the branches are large then this algorithm takes might end up in an infinite loop.

- Get the 1st link not visited from the start page
- Visit link and get 1st non-visited link
- Repeat above step till no non-visited links
- Go to next non-visited link in the previous level and repeat 2nd step

3. Page Rank Algorithm

Page rank algorithm determines the importance of the web pages by counting citations or back links to a given page [26].The page rank of a given page is calculated as

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

PR(A) Page Rank of a Website,

d damping factor

T1,….Tn links

An algorithm, taking the human factor into consideration, to introduce page belief recommendation mechanism and brought forward a balanced rank algorithm based on PageRank and Page belief recommendation which ultimately attaches importance into the subjective needs of the users; so that it can effectively avoid topic drift problems.

V. CONCLUSION

There are still various challenges that need to be overcome to effectively using semantic web. Semantic Web relies on two basic components, ontologies and semantic annotations. It relies on ontologies in order to interpret the textual content of a resource regardless of its format. Even though there have been many conceptual approximations in the field of Semantic Web in which it is assumed that resources have been semantically annotated. So, in order to take profit from the Web resources which are currently available, the extraction of features from plain text through the semantic analysis of its content and in association with the concepts of ontologies, however, building the ontology manually is an extremely complex task. Many ongoing researches are now looking at building the ontology automatically and create mapping between heterogeneous data sources This paper provides an up-to-date overview of researches that aim to enhance efficiency of SW technologies. It discussed how Semantic Web technologies evolved from the traditional ETL based to the more automatic mapping of multidimensional data.

ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of Director JSCOE, Principal JSCOE and Information Technology Department JSCOE, Hadapsar for their help in the preparation of this document.

REFERENCES

[1] Kamal Taha (2015) “RGFinder: A System for Determining Semantically Related Genes Using GO Graph Minimum Spanning Tree” IEEE Transactions On Nano Bioscience VOL.14 NO.1 pp.24-37.

[2] Lina Yao, Quan Z. Sheng,Anne. H.H. Ngu, Jian Yu, and Aviv Segev(MAY/JUNE 2015) “Unified Collaborative and Content-Based Web Service Recommendation” IEEE Transactions On Services ComputingVOL.8, NO.3 pp.453-466.

[3] Anuar, F.M.; Setchi, R.; Lai, Y.-K.(2015)“Semantic Retrieval of Trademarks Based on Conceptual Similarity”, IEEE Transactions on Systems, Man, and Cybernetics: Systems VOL., Issue: 99 pp.1-14.

[4] Huang, Yanhao; Zhou, Xiaoxin(2015)“Knowledge model for electric power big data based on ontology and semantic web” CSEE JournalVOL.1 Issue:1pp.19-27.

[5] Hai Dong and Farookh Khadeer Hussain (2014) “Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery” IEEE transaction, VOL.10 Issue:2 pp.1616-1626.

[6] SuriatiAkmal, Li-Hsing Shih, Rafael Batres(2014)“Ontology-based similarity for product information retrieval” Elsevier, Computers in Industry 65,pp.91–107

[7] H. Dong, F. Hussain, and E. Chang, O. Gervasi, D. Taniar, B. Murgante,A. Lagana, Y. Mun, and M. Gavrilova, Eds.(2009) “State of the art in semantic focused crawlers,” in Proc. ICCSA 2009, Berlin, Germany, VOL. pp.910–924.

[8] A. McCallum, K. Nigam, J. Rennie, K. Seymorey(1999) “Building Domain-Specific Search Engines with Machine Learning Techniques” AAAI Technical Report SS-99-03,pp.28-39.

[9] M. Ehrig, A. Maedche (2003) “Ontology-focused crawling of web documents” SAC’03: Proceedings of the 2003 ACM symposium on Applied computing, ACM Press, New York, NY, USA,pp.1174–1178.

[10] A. Maedche, M. Ehrig, S. Handschuh, L. Stojanovic, R. Volz (2002) “Ontology-focused crawling of documents and relational metadata” Proceedings of the11th International World Wide Web Conference WWW-2002, Hawaii.

[11] T. R. Gruber (1993) “A translation approach to portable ontology specifications” Knowledge Acquisition, VOL.5, pp.199–220.

[12] K. Graham, C. Jeremy, “Resource Description Framework (RDF): Concepts and Abstract Syntax”, http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/ , 2004.

[13] Brickley Dan, Guha RV. RDF Vocabulary Description Language 1.0: RDF Schema. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/; 2004

[14] D. Mike, G. Schreiber , “OWLWeb Ontology Language Reference”, http://www.w3.org/TR/2004/REC-owl-ref-20040210/, 2004

[15] NarasinghDeo “Graph theory with applications to engineering and computer science” PHI, 2004 Pg301

[16] Yongbin Qin and DaoyunXu “A Balanced Rank Algorithm Based on Page Rank and Page Belief recommendation”