Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 10, October 2013, pg.154 – 158

SURVEY ARTICLE

A Survey of Crawling of Untagged Web Resources Using Ontology

C.Saranya[1], Ms.R.S.Ramya[2]

[1]PG Scholar, [2]Assistant professor

[1, 2] Department of Computer Science and Engineering, Dr.N.G.P Institute of technology, Coimbatore, India

[1] me.saranngp@gmail.com; [2] ramyars.ngp@gmail.com

Abstract— The accelerated surge of the web resources is one of the current research areas. Focused crawler is one of the resources discovery systems; this is used to fetch the appropriate web pages based on the search topic from the WWW. Here, we discussed about different approaches of focused crawler for relevance prediction and also for finding the appropriate contents. Ontology is one of the techniques for describing the knowledge and for finding the relationship between the concepts. Thus, this survey deals with focused crawler using ontology.

Keywords— Ontology; Focused crawler; Onto-matching; OWL; Crawler

I. INTRODUCTION

The Semantic web is the emerging search engine for performing the metadata based search. Semantic web enables people to share and reuse of content beyond the boundaries of the application. The structure of the semantic webs is RDF, crawler, annotated/knowledge database, ontology and ranking and prioritization. The primary goal of semantic web is enable the user to find, share and combining the information more easily. The semantic web converts unstructured and semi-structured documents into “web of the data”.

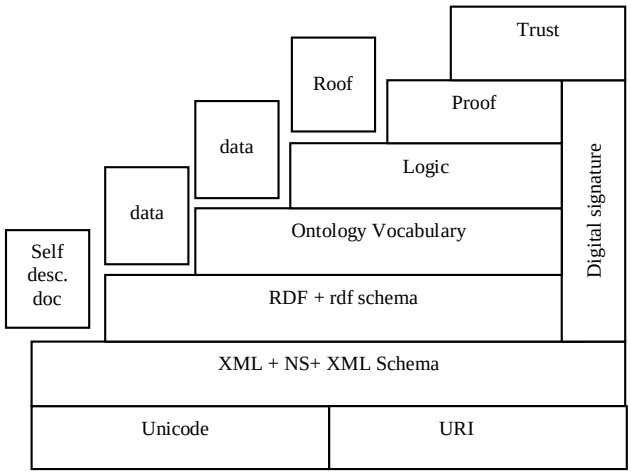

Fig. 1 Semantic web architecture [9].

Web search engine uses web crawler to update the content with other web content. Web crawler copies the entire web content that are all retrieved by the user and indexes for quick search by the user. The web crawler start searching from the starting of the seed points, here the crawler visits the seeds and identifies all the hyperlinks in that page and adds that links to the seeds, this is done recursively based on some crawler policies. The main policy of the crawler are selection policy is to download the pages, re-visiting policy is to check for changes in the pages, politeness policy is to avoid overloading websites and parallelization policy is to coordinate the distributed web crawler. The RDF is designed as a metadata data model, which is for conceptual modelling of information, is implemented in web resources. It consist of triples, they are subject, predicate and object. RDF is written in XML, so the language is called as RDF/XML.

II. ONTOLOGY

Ontology are defined as set of primitives for knowledge, it is for describing the relationship between the search topics. Commonly, ontology used in representing knowledge, retrieval of information, understanding of natural language and web services [2]. Ontology is for determining the relation among the various concepts and used to find the distance between the pair of concepts [1].

Ontologies are related to each other through ontology mapping, which is challenging factor. Ontology mapping refers to finding relation between the different ontologies. The major goal of ontology is to share the understanding of information, to analyse domain knowledge, domain assumption, domain knowledge reuse and separation of domain knowledge from operational knowledge [2]. Owl is the ontology specification language. OWL is used to explicitly represent the meaning of the terms and their relationship between those terms. Owl contains three sublanguages [2].

- OWL Lite: it is to support classification for the users and also for simple constraints. It has less formal complexities than the Owl DL.
- OWL DL: while retaining the computational completeness and decidability, OWL DL supports the users for maximum expressiveness. This includes OWL language constructs and it is used under only certain constraints.
- OWL Full: it support users with no computational expressiveness completeness for achieving maximum expressiveness and syntactic freedom with RDF. It uses the OWL primitives and also RDF and its Schema with it.

III. FOCUSED CRAWLER

Focused crawler is used to retrieve the relevant web pages from the bookmarked site based on the search topic. Focused crawler analyses the boundary of crawl, to find the most relevant pages for crawl and also the irrelevant pages by using the links. To achieve the goal, proposed two hypertext mining program (i) one is to evaluate the relevance between the web resources is called classifier. (ii) Next is to identify the nodes which are more relevant to the pages is called distiller. The focused crawler performance is mainly depends on the abundance of the link, these link is based on the search topic given. Focused crawler start searching from the root set is to acquire the relevant pages; it does not loses its way for searching and is robust [3]. Focused crawler classified into:

1. Classic focused crawler [1]: this utilizes the knowledge about the search topic; by using it determines the relevance between the web pages. It downloads the higher priority link by search the interested pattern, these downloaded priority is based on the search topic and the link information to that search topic. The model computed for classic focused crawler is vector space model [5].
2. Learning focused crawler [1]: this crawler uses the crawling guidelines and also follows the pre-defined guidelines for assigning the visit priorities. Those guidelines are based on the updated training set in it. For building training set, learned focused crawler need more processing time. The model for computing it is context graph [6] and Hidden markov model [7].
3. Semantic crawler [4]: this is one of the modifications of the classic focused crawler. For retrieving the relevant web pages, it uses social web and the semantic knowledge.

IV. ONTO-MATCHING

Ontology specifies the entities of interest in domain in terms of attributes for classes which defines the concepts in ontologies , individuals is the object instances, properties which finds the possible associations between the individuals and data types are the values can have with it. Onto-matching is to derive the adjustment between the ontologies. Ontology-matching is also known as onto-match or onto-matching or ontology-map. Onto-matching defines the semantic relationship between the concepts in different ontologies by using many formal languages like OIL, DAML+OIL and RDF[8]. Ontology deals the problem of mapping pattern, when matching the element with same meaning between different ontologies it reflect the internal structure of the mapping pattern. Consider two different ontologies l and m with entities l' and m'. By using the entities find the coherence between the elements based on the same meaning and also find the equivalence between the elements [2].

V. RELATED WORK

This section deals with the discussion about the focused crawler using ontology. In first method maryam hazman [10], gives the survey about the focused crawler problem which are faced during the search of relevant web pages. This deals with the ontology based focused crawler, structure based focused crawler and other focused crawler approaches. The ontology based focused crawler is proposed to identify the relevant web pages, before those pages get downloaded and processed. Structure based focused crawler is to classify the relevance between the web pages by using the structure of the web pages.

In the next method, Prasant Singh Yadav, Mala Kalra, K.P Yadav [11], proposed the ontology based content based focused crawler this deals with the crawl boundary links, those are most likely to be relevant topic based on the search interest. This technique searches in the ontology instead of searching in the web db. For every short interval of time, the ontology periodically updates its content. This technique displays only the related content for user needs; it doesn't display all the information which is all not needed for the user. This technique achieved two benefits this reduces the number of the results invoked and another one it reduces the time by rejecting the irrelevant information during search which makes the search easier for the user.

In this method, Deepali Dev [12], this paper deals with the different relevance prediction technique comparison for the focused crawler based on content i.e., fish search and shark search relevance prediction based on the link analysis. In fish search this start searching from the set of seed points based on the topic given by the user. In this technique matches the content based on the query and then find their neighbourhood, so this system search is query driven. This search algorithm is like directed graph, it treats web pages as nodes and the links between those web pages as edges, the searching is done like traversing through the edges of the directed graph. It uses binary evaluation for finding the relevancies; it uses 1 for relevant pages and 0 for irrelevant pages. It stores URL in the list and it prioritizes it. There is one problem in this search is the low differentiation of the page priority in lists. Next search algorithm is the modification of the fish search is the shark search, it differing from fish by two ways, first it inherits the modified value of the score of its parent and next, that score is combined with the score of the anchor value of the links. This technique uses fuzzy score instead of the binary evaluation for relevance and it also uses the vector space model for finding the relevance's. the fuzzy score contains the score values from 0 to 1, 0 indicates there is no similarity between the pages and 1 indicates the perfect match i.e., relevant pages.

Relevance prediction based on content and link analysis uses the HAWK algorithm. This algorithm predicts the relevant web pages, selects and prioritizes its URL. This algorithm not only uses the web content along with it uses link information for predicting relevancies.

In this method, Ignacio Garcıa Dorado [13], this paper gives the brief survey about the algorithms of focused crawler. In this we deal with breadth-first, best-first search algorithm Graphic Context Algorithm. The breadth-first algorithm is implemented with depth-first search as FIFO. Here, the crawler start searching from the relevant web pages and it finds the next relevant web pages that are most likely to be close to relevant web page, which is fetched by using the minimum number of hips to the root.

The algorithm follows [13]:

insert in queue(seeds)
while more links in queue do
link := dequeue first inserted
doc := fetch(link)
out links := extract links(doc)
insert in queue last pos(out links)
end while

In best-first search algorithm, it selects the best links by using the greatest rank/scores; the rule is applied to fetch the select link from the frontier. Naïve Bayes and Cosine similarities will be used for classifier algorithm and for scoring uses SVM and string matching. The algorithm follows [13]:

insert in ready queue(seeds)
while true do
if more links in ready queue then
link := dequeue best
doc := fetch(link)
score := apply rule(doc)
out links := extract links(doc)
save score(out links, score)
else
sorted links := sort(non processed queue)
insert in ready queue(sorted links)
end if
end while

In Graphic Context Algorithm, for retrieving the better pages, first it manually selects the relevant pages in first layer, by using that layer it fetches the next nearest relevant pages as next layer the fetched pages are stored in the queue which is corresponding to that layer. Each layer contains the corresponding queue; it stores all the fetched pages in that layer. Again, it finds for the next layer, this processes continues till the whole process completes. For obtaining the context graph it must attain whole process. The deep-seated layer in the circle will attain the better relevant pages. The algorithm follows as [13]:

insert in ready queue(seeds)
while true do
if more links in ready queue then
link := dequeue best
doc := fetch(link)
for i = 0 to num layers do
scoreL[i] := classifyLi(doc)
end for
layer belong := maximum(scoreL[i])
out links := extract links(doc)
save score(out links, scoreL, layer belong)
else
sorted links := sort(non processed queue)
insert in ready queue(sorted links)
end if
end while

In anchor algorithm, it combines the parent pages with the anchor text, so it yields better results than the classifying through the pages alone. In history path algorithm, for finding the better relevant pages it uses metric for finding the distance between the relevant pages with the next relevant pages and uses that distance to find next relevant pages [13].

In this method, Yves R. Jean-Mary et. al [8], this paper deals with the novel algorithm which is named as the Automated Semantic Matching of Ontologies with Verification (ASMOV). To calculate the similarities between the two ontologies it

derives the arrangement and verifies it to assure there is no deviations, it uses lexical and structural characteristics. This paper gives high accurate results by finding the similarities between two ontologies with the process semantic verification.

In the method, Ondrej Svab [14], in this paper deals with the ontology mapping. For finding the difference between the two ontology need to consider mapping pattern with it. The mapping pattern is in form of graph structure and it deals with the internal structure of the ontologies. In graph structure, the classes, properties are represented as nodes and the mapping between classes, relation between classes represents the edges. This paper have two concerns for ontology design patterns: naming conventions and structural patterns.

In this method, Dominique Ritze, Christian Meilicke, Ondrej Svab-Zamazal, and Heiner Stuckenschmidt [15], this paper is to detect the simple coherence among the atomic concepts and their properties. States of the ontology matching techniques are used for limited in it. This approach doesn't await on machine learning technique, it exploits the input arrangements it consists of the simple coherences. In many cases, it fails to generate highly precise alignments.

VI. CONCLUSIONS

As this survey shows, the different focused crawler algorithms, ontology and onto mapping in the semantic web. This work is useful for researcher for finding the relevance prediction and also for finding the exact result, the extraction of untagged web pages from web is done by using any one of the above technique with better relevant web pages and also for finding the similarities between the ontologies by using any one of the mapping pattern. This is useful for retrieving the untagged web resources by focused crawler technique using ontologies

REFERENCES

[1] Punam Bedi , Anjali Thukral, Hema Banati, "*Focused crawling of tagged web resources using ontology"*, Computers and Electrical Engineering 39 (2013) pp 613–628.

[2] Ramya.R.S, Raja Ranganathan.S, Dr. S.Karthik, *"A Brief Survey on Improving the Efficiency of Revisiting Concepts in Semantic web"*, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 1, ISSN: 2278 – 7798 , January 2013.

[3] Chakrabarti S, van den Berg M, Dom B. "*Focused crawling: a new approach to topic-specific web resource discovery"*. Comput Netw 1999;31(11–16):1623–40.

[4] Batsakisa S, Petrakisa EGM, Milios E. "*Improving the performance of focused web crawlers"*. Data Knowl Eng 2009;68(10):1001–13.

[5] G. Salton, A. Wong, C.S. Yang, *"A vector space model for automatic indexing"*, Communications of the ACM 18 (11) (1975) 613–620.

[6] M. Diligenti, F. Coetzee, S. Lawrence, C. Giles, M. Gori, "*Focused crawling using context graphs"*, Proceedings of the 26th International Conference on Very Large Databases (VLDB 2000), 2000, pp. 527–534.

[7] H. Liu, J. Janssen, E. Milios, *"Using HMM to learn user browsing patterns for focused web crawling"*, Data & Knowledge Engineering 59 (2) (2006) 270–329.

[8] Yves R,. Jean-Marya E, Patrick Shironoshitaa, Mansur R. Kabuk, "*Ontology Matching with Semantic Verification*", Web Semantics: Science, Services and Agents on the WWW, vol.21, 2009.

[9] http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html.

[10] Maryam Hazman *"A Survey Of Focused Crawler Approaches"*, Journal of Global Research in Computer Science Volume 3, No. 4, April 2012 ISSD-2229-371X.

[11] Prasant Singh Yadav, Mrs Mala Kalra, Dr. K.P Yadav *"Enhancing the performance of web Focused CRAWLer using ontology"*, International Journal of Computers &Technology Volume 4 No. 2, March-April, 2013, ISSN 2277-3061.

[12] Deepali Dev *"Relevance Prediction In Focused Crawling: A Survey"*, Journal Of Information, Knowledge And Research In Computer Engineering Issn: 0975 – 6760| Nov 12 To Oct 13 | Volume – 02, Issue – 02

[13] Ignacio Garc´ıa Dorado "*Focused Crawling: algorithm survey and new approaches with a manual analysis"*, Department of Electrical and Information Technology Lund University, April 7, 2008.

[14] Ondrej Svab, "*Exploiting Patterns in Ontology Mapping*", 6th International Conference on Semantic Web Conference, 2008.

[15] Dominique Ritze, Christian Meilicke, Ondˇrej Svab-Zamazal, and Heiner Stuckenschmidt "*A pattern-based ontology matching approach for detecting complex correspondences, University of Mannheim"*, dritze@mail.uni-mannheim.de, fchristian, heinerg@informatik.uni-mannheim.de2 University of Economics, Prague, ondrej.zamazal@vse.c.

