International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in; editorijoer@gmail.com

Vol.5., Issue.1., 2017
January-February

RESEARCH ARTICLE

ISSN: 2321-7758

CONTEXT TREE BASED SEARCHING FOR OPTIMIZATION USED AS ADVANCE TECHNOLOGY IN WEB CRAWLERS

PALAK[1], VIKAS SRIVASTAVA[2]
[1]M.Tech Scholar Computer Science Engineering, IIMT Engineering College Meerut, Uttar Pradesh
Email:palakagarwal1508@gmail.com
[2]Supervisor, IIMT Engineering College Meerut, Uttar Pradesh
Email:vikassrivastava@gmail.com

PALAK AGARWAL

ABSTRACT

Data Mining techniques have moved to a new level due to exponential growth of data and users query types in knowledge management systems. As knowledge is an essential and volatile asset and web is the only available hub from where it can be retrieved, most of the search engine techniques focuses on keyword searching and simulating existing results. To achieve more prominent results and relevant URLs for the users query, the problems of crawlers (polysemy and synonymy) must be removed. In this paper, we have focused on removing these existing problems by introducing the concept of adaptive web searching based on structural ontology. Contexts are related with each other on the basis of properties and meta-keywords. Further, simulation of architecture is also shown which provides promising results.

I. INTRODUCTION

Difficulties of search engines are increasing day by day due to huge size and multi-disciplinary nature of Internet which keeps on growing its knowledge, which is again an unstructured and volatile asset. Issues related to inference, relate or represent knowledge in multiple formats is a main issue. Also, locating useful information over entire internet which must be best among all available information according to users requirements. To intelligently fetch the most relevant results many models had been proposed by various researchers, however as the demand of the users keeps changing, the ways to extract desired information should be changed to meet such demands. Basically, networks bandwidth gets waste because a major portion of retrieved information is irrelevant. Simple crawlers are not much focused due to their inefficiency in solving major problems of information retrieval, namely polysemy (one word, many meanings) and synonymy(many words, same meaning). Along with it, locating precise resources according to relevancy in a prescribed amount of time is a huge challenge for all purpose single purpose crawlers. Now-a-days, many documents contains desired semantic information, even though they do not contain user-specified keywords. This context, or sematic of the document can be easily derived using the concept of Ontology. Ontology defines entities, relations, properties and keywords that fundamentally exists for a particular domain. Also, we can express formal specification of terms in domain and relations among them. Well use ontology, to store the relationship between the words and contexts to represent this knowledge in the form of network (inter-connection of nodes and edges, where nodes will represent the word and edges will re represent the typeof relationship between them). Ontologies are used across a number of domains. Ontologies often contain a model of domain, its taxonomy and relation- ship between its entities.

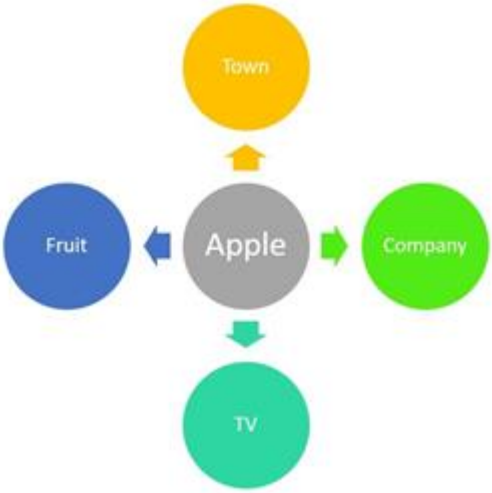


Fig. 1.A simple ontology for word apple consisting of multiple contexts

II. RELATED WORK

The idea of implementing ontologies in the field of focused searching is fresh and to the best of our knowledge until now it has not been pursued in research. However, there are related works scattered around the world that share some similarities with our approach.

A. Highly efficient architecture for scalable focused crawling using incremental parallel web crawler, P. Jaganathan, T. Karthikeyan, [1]

In this paper concurrent crawling of web pages related to multiple pre-defined topics is conducted which solves the issue of URL distribution. A compound decision model is developed which makes the decision making process multi objective by considering various factors of searching like load balance and relevance concurrently. Also, updating frequency is also handled by local repository decision.

B. Ontology based Context Synonymy Web Searching Eakansh Manglik, Priyanka Sharma, Paramjeet Rawat, Nidhi Tyagi, [2]

In this paper a keyword matching technique is applied which fetch the results from web repository. A structural approach is followed which focuses on unstructured knowledge and solves the issues of polysemy and synonymy by constructing conceptual ontology for each domain and their contexts. A simulation architecture is implemented which shows promising results.

C. Context based Web Indexing for Storage of Relevant Web Pages Nidhi Tyagi, Rahul Rishi and R.P. Aggarwal, [3]

This paper proposes a technique for indexing the keyword extracted from the web documents along with their contexts wherein it uses a height balanced binary search (AVL) tree, for indexing purpose to enhance the performance of the retrieval system. This data structure is able to support dynamic in- dexing, which is especially important for environments where documents are changed frequently. If the planning about the arrangement of the keywords is done then AVL tree can be achieved.

D. Concept based Focused Crawling using Ontology, S.Thenmalar and T. V. Geetha, [4]

The paper proposes and extends the idea of focused crawling by prioritizing the queue of URLs downloaded by the focused crawler. This idea is supported by designing a conceptualized vector which is obtained by combining concept vectors of individual pages associated with seed URLs. The conceptual rank is based on comparison between conceptual vectors at each depth, across depths and between the overall topics indicating seed concept vector.

E. Contextual Ontology: A Storage Tool for Extracting Con- text from Web Pages, Nidhi Tyagi, Rahul Rishi and R.P. Aggarwal, [5]

Contextual ontology helps in the knowledge-full indexing of documents, providing semantic structure to the document. The advantage of such model is sharing common understanding of the structure of context information among users, devices and services to enable semantic interoperability. It also enables reuse of domain knowledge, i.e., building a large ontology by integrating several ontologies describing portions of the larger domain. Also, it enables formal analysis of domain knowledge, for example, context reasoning becomes possible by clearly defining context ontology.

III. PROPOSED WORK

The major problem in the existing search methodology is that, almost all web search engines work on Keyword Match- ing Technique and are trapped in the two major problems of information retrieval i.e. polysemy and synonymy. To solve the above problem, we break the traditional searching into a two-step process:

A. Architecture of Ontology Based Adaptive Web-Searching

The architecture of the proposed work is divided into two phases i.e. firstly, Architecture of Ontology and secondly, Architecture of Searching.

1) Architecture of Ontology: It defines how the ontology is to be developed and stored/update timely in the base, for future references

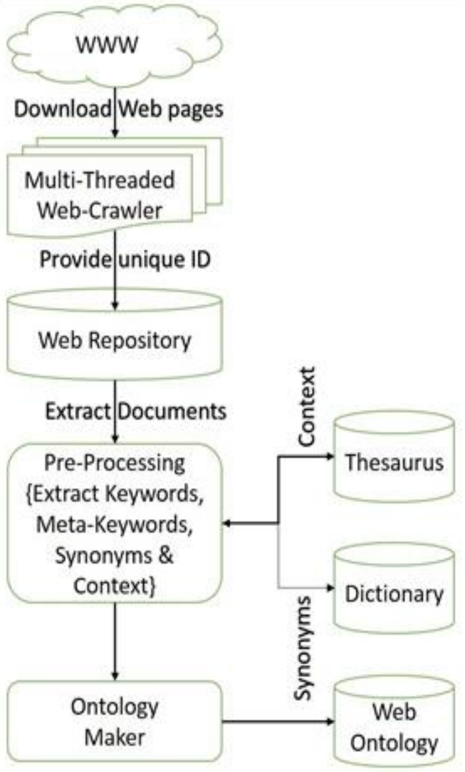


Fig. 2. Architecture of Ontology Construction

It shows mainly the following steps:

Step 1: Web pages are downloaded from the Internet by the multi-threaded crawler. An unique ID is provided to every web page for identification. All the downloaded web pages are stored in a base known as Web-Repository.

Step 2: From web repository, downloaded documents are ex- tracted for preprocessing in which keywords, meta-keywords, synonyms and contexts are extracted.

Step 3: For extracting multiple contexts of any word (ifany), thesaurus is utilized and for finding synonyms of any word, dictionary is utilized.

Step 4: Once all the required information is gathered, i.e. keywords, their contexts, related words and properties, for any particular word, its ontology is created and is stored in web ontology afterwards.

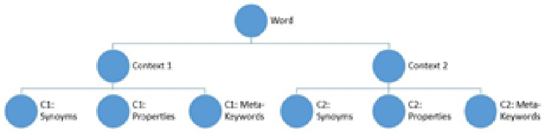


Fig. 3. Diagrammatic representation of ontology stored at backend.

2) **Architecture of searching**: Architecture of searching: It defines how the searching process is to be performed, what all preprocessing, post processing and references are required in the searching at different levels.

Step 1: For each query provided by the user, the architecture extracts keywords and removes stop-words, if any from that query to make the search more relevant.

Step 2: A condition is checked for each query to determine whether the query contains multiple keywords or not.

Step 2.a (If multiple keywords are there): In this step, a processing is done to determine the relationship between keywords so that the appropriate context of the query is decided. This process takes the help of Web-Ontology.

Step 2.b (If multiple keywords are not there): In this step, the keyword is further processed to retrieve the desired context of the word with the help of web ontology.

Step 3.a (If multiple contexts are available): All of the available contexts are loaded on the user screen and user selects the desired context from the list.

Step 3.b (If multiple contexts are not available): The entire qualifying URL list based on context criteria is retrieved from web repository.

Step 4: As the user select any context, the corresponding ontology is retrieved from the context synonymy ontology and the qualifying URL list based on ontology criteria, is retrieved from web repository.

Step 5: Ranking of URLs is done on the basis of their relevance ratio

Step 6: Finally sorted and focused results are displayed to the user.

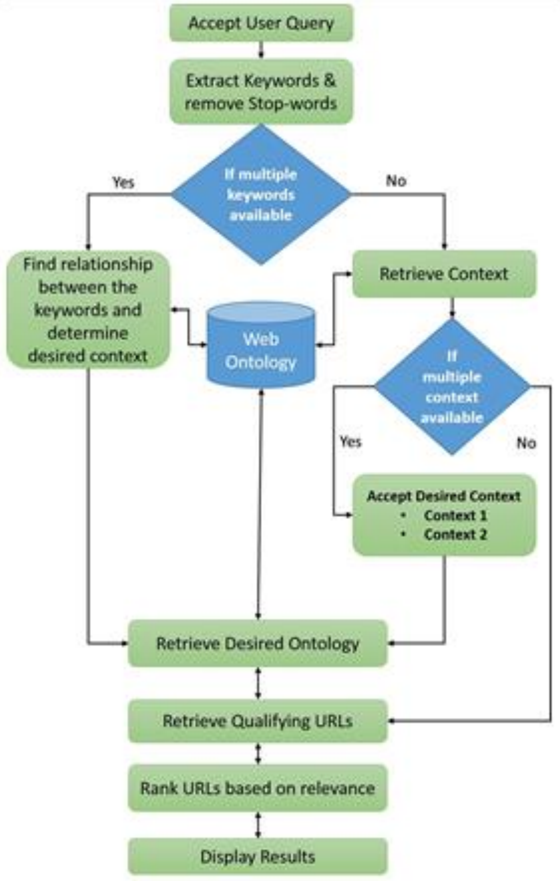
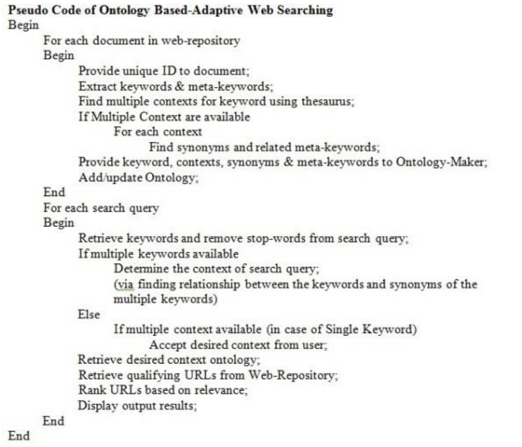

Accept User Query
Extract Keywords & remove Stop-words
If multiple keywords available
Yes
No
Find relationship between the keywords and determine desired context
Web Ontology
Retrieve Context
If multiple context available
Yes
No
Accept Desired Context
• Context 1
• Context 2
Retrieve Desired Ontology
Retrieve Qualifying URLs
Rank URLs based on relevance
Display Results

Fig. 4. Architecture of Searching

The above diagram depicts the architecture of searching. It consists of following steps:

IV. PSEUDO CODE OF ONTOLOGY BASED-ADAPTIVE WEB SEARCHING



```
Pseudo Code of Ontology Based-Adaptive Web Searching
Begin
    For each document in web-repository
    Begin
        Provide unique ID to document;
        Extract keywords & meta-keywords;
        Find multiple contexts for keyword using thesaurus;
        If Multiple Context are available
            For each context
                Find synonyms and related meta-keywords;
        Provide keyword, contexts, synonyms & meta-keywords to Ontology-Maker;
        Add/update Ontology;
    End
    For each search query
    Begin
        Retrieve keywords and remove stop-words from search query;
        If multiple keywords available
            Determine the context of search query;
            (via finding relationship between the keywords and synonyms of the
            multiple keywords)
        Else
            If multiple context available (in case of Single Keyword)
                Accept desired context from user;
        Retrieve desired context ontology;
        Retrieve qualifying URLs from Web-Repository;
        Rank URLs based on relevance;
        Display output results;
    End
End
```

Fig. 5. Pseudo Code of Ontology Based-Adaptive Web Searching.

V. SIMULATION

Ontology based adaptive web searching is a simulated search engine which uses ontology to support searching and provide a structure to unstructured web repository. Googles web repository is utilized as its web page repository and the- saurus.com as knowledge base for building ontology. Googleprovide Search API [9] (Application Programming Interface) and thesaurus.com provide dictionary API. The model On- tology based Adaptive Web Searching is implemented using Microsoft Dot.Net. The frontend is designed in Asp.Net, the scripting is done in C-hash and ontology is developed using XML [6] as a Knowledge Base support to the complete system.

A. Snapshots of Simulated Model

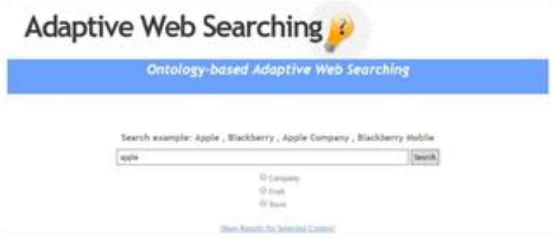
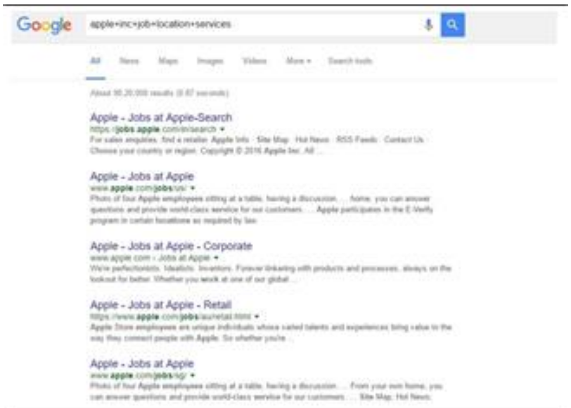

Adaptive Web Searching
Ontology-based Adaptive Web Searching
Search example: Apple , Blackberry , Apple Company , Blackberry Mobile
apple
Search
○ Company
○ Fruit
○ Rock
Show Results for Selected Context

Fig. 6. Query page of the simulated Search Engine for word Apple + context Inc.



apple+inc+job+location+services

About 90,20,000 results (0.87 seconds)

Apple - Jobs at Apple-Search
https://jobs.apple.com/in/search ▾
For sales enquiries, find a retailer. Apple Info · Site Map · Hot News · RSS Feeds · Contact Us · Choose your country or region. Copyright © 2016 Apple Inc. All ...

Apple - Jobs at Apple
www.apple.com/jobs/us/ ▾
Photo of four Apple employees sitting at a table, having a discussion. ... home, you can answer questions and provide world-class service for our customers. ... Apple participates in the E-Verify program in certain locations as required by law.

Apple - Jobs at Apple - Corporate
www.apple.com › Jobs at Apple ▾
We're perfectionists. Idealists. Inventors. Forever tinkering with products and processes, always on the lookout for better. Whether you work at one of our global ...

Apple - Jobs at Apple - Retail
https://www.apple.com/jobs/au/retail.html ▾
Apple Store employees are unique individuals whose varied talents and experiences bring value to the way they connect people with Apple. So whether you're ...

Apple - Jobs at Apple
www.apple.com/jobs/sg/ ▾
Photo of four Apple employees sitting at a table, having a discussion. ... From your own home, you can answer questions and provide world-class service for our customers. ... Site Map · Hot News.

Fig. 7. Search results for the word Apple and the selected context Inc by Ontology based Adaptive Web searching

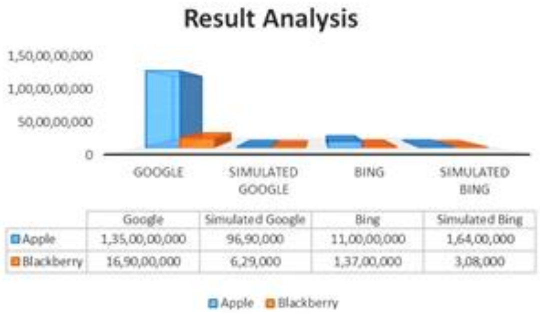
VI. RESULT ANALYSIS

We have considered two words (i.e. Apple, Black berry) ini- tially. Both of the words have multiple contexts available. The above bar graph shows the result analysis for these two words (i.e. Apple, Black berry) over three leading search engines (i.e. Google, Yahoo and Bing) and one of our implementation (Ontology based Adaptive Web-Searching).

Above chart shows, different search engines with three words on x-axis and number of output results on y-axis. We have analyzed the same query over various search engines and the results are displayed in the form of bar-graph. It has been observed that in comparison to other search engines the results are more specific and less in number. It has been concluded by viewing the graph that the search becomes more focused.

VII. CONCLUSION

Adaptive representation of the ontology plays an important role in supporting the task of document classification and iden- tification. The proposed technique improves the performance of the searching system in terms of accuracy and efficiency for retrieving more, appropriate documents as per the users requirements. As the documents are stored contextually the relevant URLs are made available to the user in short span of time. Also it uses available contextual information and ontologies to rank the underlying documents as well as the search results. The use of contextual information results in better ranking of the documents and hence results in higher quality of the retrieved results. The use of multiple keywords for making relational query makes the searching process accurate and focused.



	Google	Simulated Google	Bing	Simulated Bing
Apple	1,35,00,00,000	96,90,000	11,00,00,000	1,64,00,000
Blackberry	16,90,00,000	6,29,000	1,37,00,000	3,08,000

Fig. 8. Result Analysis

VIII. FUTURE WORK

We can further extend the architecture of the ontology by using specific learnings like Supervised and Unsupervised which has refined properties to define any word in a much more specific level and helps in making the search more focused and effective. The idea of pruning the ontology by selective filtering of contexts by the crawler makes the search more generic and conceptualized in a unique way.

REFERENCES

[1]. P. Jaganathan, T. Karthikeyan, Highly efficient architecture for scalable focused crawling using incremental parallel web crawler, Journal of Computer Science, 2014

[2]. Eakansh Manglik, Priyanka Sharma, Paramjeet Rawat, Nidhi Tyagi, Ontology based context synonymy web searching, Information systems and Computer Networks, 2013.

[3]. Nidhi Tyagi, Rahul Rishi and R.P. Aggarwal Context based Web Indexing for Storage of Relevant Web Pages, International Journal of Computer Applications, Volume 40 No.3, (Page No. 1-5), 2012.

[4]. S.Thenmalar and T. V. Geetha Concept based Focused Crawling using Ontology, International Journal of Computer Applications, Volume 26 No.7, (Page No. 29 32), 2011.

[5]. Nidhi Tyagi, Rahul Rishi and R.P. Aggarwal, Contextual Ontology: A Storage Tool for Extracting Context from Web Pages,

International Journal of Computer Applications, Volume 56 No.7, (Page No. 30 34), 2012.

[6]. Learn XML http://www.w3schools.com/xml.

[7]. www.coursera.com

[8]. Client based architechture http://en.wikipedia.org.

[9]. Google custom search API , http://developers.google.com