ISSN:0975-9646

A Study of Focused Web Crawlers for Semantic Web

Nidhi Jain[1], Paramjeet Rawat[2]

[1]*Computer Science And Engineering, Mahamaya Technical University*
Noida, India

[2]*IIMT Engineering College*
Meerut, India

***Abstract*— Finding useful information from the web which has a large and distributed structure requires efficient search strategies. Focused crawlers selectively retrieve Web documents that are relevant to a predefined set of topics. To intelligently make decisions about relevant URLs and web pages, different authors had proposed different strategies. In this paper we review and compare focused crawling strategies, studied and published during the past few years. We also give the details of various issues related to focused crawling.**

***Keywords*— Semantic web, crawler, focused, ontology.**

I. INTRODUCTION

SEMANTIC WEB: The semantic web is a vision of information that can be readily interpreted by machines, so machines can perform more of the tedious work involved in finding, combining, and acting upon information on the web. Semantic Web technology is to address the problem by structuring the content of the web and extract maximum benefit from the processing power of machines and existing web. The Semantic Web, as originally envisioned, is a system that enables machines to "understand" and respond to complex human requests based on their meaning. Such an "understanding" requires that the relevant information sources be semantically structured. Semantic Web technology aims to provide meaning to web contents. A query posed by a user may require information retrieval from a number of web sites. These web sites advertise their capabilities using Web Services. For automatic retrieval and processing of information from several sources, software agents perform the orchestration of web services on the basis of user defined preference parameters.

There are various types of crawlers out of which focused crawlers are popularly used.

FOCUSED CRAWLER: Focused crawler is used to collect those web pages that are relevant to a particular topic while filtering out the irrelevant. Thus focused crawling can be used to generate data for an individual user. There are three major challenges for focused crawling:

(i) It needs to determine the relevance of a retrieved web page.

(ii) Predict and identify potential URLs that can lead to relevant pages.

(iii) Rank and order the relevant URLs so the crawler knows exactly what to follow next.

It is attractive only for certain domain and not for the whole web. These search engines or companies tries to gain information in their field of activity through focused crawler. A focused crawling algorithm loads a page and extracts the links. By rating the links based on keywords the crawler decides which page to retrieve next. The Web is traversed link by link and the existing work is extended in the area of focused document crawling. For this we not only use keywords for the crawl, but also depend on high-level background knowledge with concepts and relations, which are compared with the texts of the searched page. This is how a direct focus can be achieved. There are various categories in focused crawlers:

(a) Classic focused crawler
(b) Semantic crawler
(c) Learning crawler

(a)Classic focused crawlers [9] guide the search towards interested pages by taking the user query which describes the topic as input. They assign priorities to the links based on the topic of query and the pages with high priority are downloaded first. These priorities are computed on the basis of similarity between the topic and the page containing the links. Text similarity is computed using an information similarity model such as the Boolean or the Vector Space Model (VSM) [10].

(b) Semantic crawlers [11] are a variation of classic focused crawlers. To compute topic to page relevance downloaded priorities are assigned to pages by applying semantic similarity criteria: the sharing of conceptually similar terms defines the relevance of a page and the topic.Ontology is used to define the conceptual similarity between the terms [11–12].

(c)Learning crawlers [13] uses a training process to guide the crawling process and to assign visit priorities to web pages. A learning crawler supplies a training set which consist of relevant and not relevant Web pages in order to train the learning crawler [13,14]. Links are extracted from web pages by assigning the higher visit priorities to classify relevant topic. Methods based on context graphs [15] and Hidden Markov Models (HMM) [16] take into account not only the page content but also the link structure of the Web and the probability that a given page (which may be not relevant to the topic) will lead to a relevant page . There are some issues related to semantic web which are discussed below.

II. ISSUES RELATED TO SEMANTIC WEB CRAWLER

The issues related to semantic web crawler are:

A. *Input:* Number of starting (seed) URLs and (in the case of focused crawlers) the topic descriptions are inputted into crawlers. It can be the description of a list of keywords for classic and semantic focused crawlers.

B. *Page downloading:* Extracted pages of the downloaded links are placed in a queue.Queue entries are reordered in a focused crawler by applying content relevance or importance criteria or links may be excluded for further expansion(generic crawlers may also apply importance criteria to determine pages that are worth crawling and indexing).

C. *Content processing:* Downloaded pages are lexically analyzed and reduced into term vectors.According to VSM each term vector is denoted by its term frequency-inverse frequency vector (tf-idf). Here we used precompiled idf weights, provided by the IntelliSearch6Web search engine holding idf statistics for English terms.

D. *Priority assignment:* Extracted URLs from downloaded pages are placed in a priority queue where priorities are considered based on the type of crawler and user preferences. It can vary from simple criteria to more involved criteria (e.g. criteria determined by a learning process) i.e page importance or relevance to query topic (computed by matching the query with page or anchor text)

E. *Expansion:* URLs are selected for further expansion and steps (b)–(e) are repeated until some criteria (e.g. the desired number of pages have been downloaded) are satisfied or system resources are exhausted..

III. FOCUSED CRAWLING STRETIGIES

A. Special Purpose Approach

The M. Hersovici et al.[20] proposed the “shark search” algorithm which is the refined version of “fish search”. A dynamic Web site mapping enables users to tailor Web maps to their interests by representing the shark search. The advantage of this approach that it is more significant than fish search algorithm. The author et al.[21] proposed an improved quality of web navigation through effective focused crawling. The focused crawler generates meta data and resultant pages based on which the priority of the extracted links is calculated which helps in checking the similarity of web pages of the keyword. It is used to improve the coverage of specific topic by traversing the irrelevant pages. The disadvantage of this approach is it is time consuming in crawling the web pages but it has better performance than BFS crawler. Y. Zhang et al.[22] proposed an improved Page Rank algorithm called as "To-Page Rank", and present a crawling strategy which is used to combine the topic similarity of the hyperlink metadata. It has better performance than the Breath-first and Page Rank algorithms.

B. Structure Based Approach

Jon M. Kleinberg[19] proposes notion of authority on the basis of algorithmic formulation, based on the relationship between a set of relevant authoritative pages and the set of hub pages that join them together in the link structure. This formulation is connected to the eigenvectors of certain matrices associated with the link graph which motivates additional heuristics for link-based analysis. The problem in the previous crawlers was that they took longer time to crawl the relevant pages and the searching quality was not appropriate. This approach provide effective search methods in which one needs a way to filter a small set of authoritative pages from the huge collection of relevant pages. The main obstacle which one faces is how to define whether it is authoritative or not. In this paper [18], Google is designed to crawl and index the Web efficiently which provides an in-detail description of our large-scale Web search engine and addresses the question of how to build a practical large-scale system which can exploit the additional information present in hypertext. It also deals with the problem of uncontrolled hypertext collections where anyone can publish anything they want. The commercial search engines had focused more on efficiency rather than searching quality of URLs, while this approach focuses more on the quality of search to decide what old pages should be recrawled and what new ones should be crawled.

C. Block Partitioning Approach

In the paper[5] author proposed block partitioning technique in which blocks are partitioned by VIPS algorithm and calculating the sum of all block relevancy score in one page and then calculate the URL score to identify whether URL is relevant or not for specific topic. The previous crawlers focused on measuring the relevancy of a page and calculate the URL’s score based on whole page’s content and these web pages contains multiple topics which may or may not be related to the given domain. The evaluation of the whole page may cause a lot of irrelevant link because it contains noises due to which the evaluation of the relevant information is ignored. The advantage of this approach is that it calculates the score based on the relevancy of content blocks of web page rather than calculating the URL score of that URL which contains the irrelevant links.Y.Sun et al. [28] proposed a new hybrid approach to focused crawling based on meta search algorithm to achieve a wider crawling range and VIPS algorithm for the relevance computation of a web page to partition into set of blocks of a web page which reflects the semantic structure of the page. Another hybrid focused crawler was proposed by Mohsen Jamali et al.[30] that uses link structure of documents and similarity of pages to crawl the web .In experimental evaluations, the crawler is compared with the unfocused one. For this their behavior is studied in two different tests, one uses a hub as seed page, and the other uses a non-related page. It shows superiority over non-focused one with a high harvest rate. The advantage of this crawler is that its harvest rate is better than a BFS crawler but the time consumption is more than an ordinary crawler.

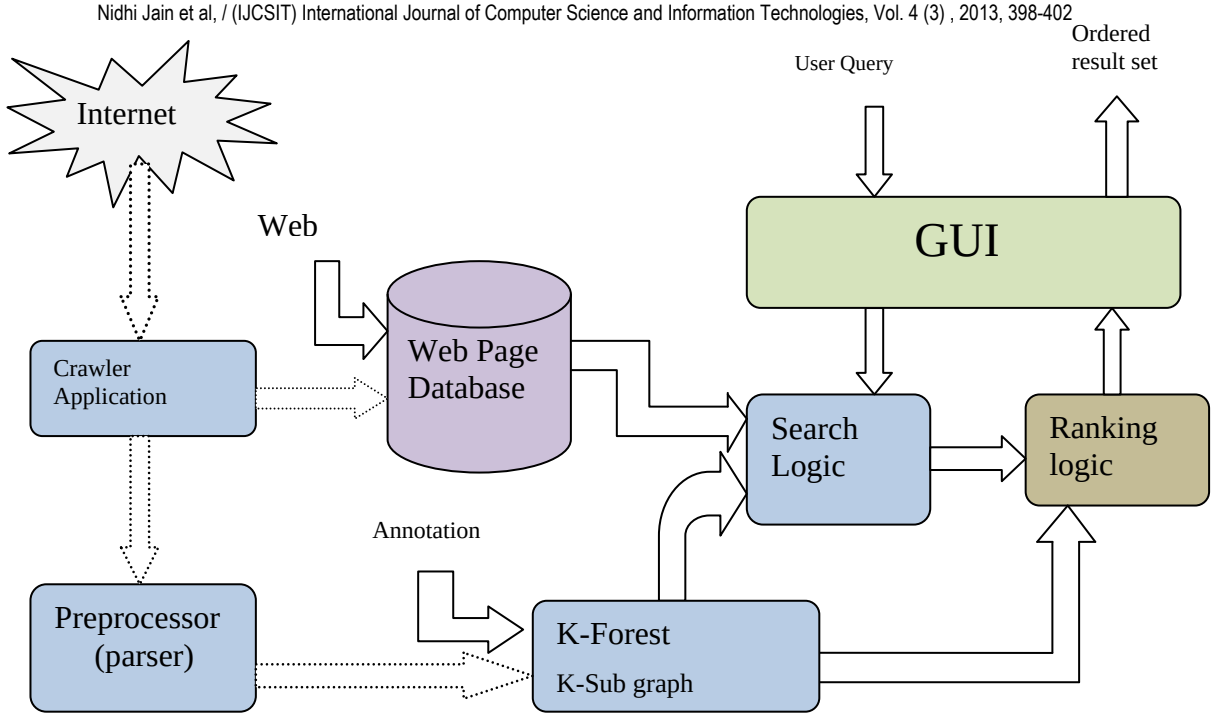

Fig 1: Architecture Of Semantic Web

D. Fuzzy Go-Based Approach

Few author proposes fuzzy based approaches.Jalilian et al.[7] addressed a Fuzzy-Go based search engine.First, a fuzzy ontology based on the concept of fuzzy logic was used to capture the similarities of terms in the ontology that offers appropriate semantic distances between terms to accomplish the semantic search of keywords. Thus it can automatically retrieve the web pages which contain keywords of similar terms. After that the domain classification of web pages offers users to select the appropriate domain for searching ,that excludes web pages in the inappropriate domains to reduce the search space and to improve the search results. The advantage of this crawler is that it has higher precision rate compared with breadth-first and best-first search crawlers which shows that it has higher efficiency in terms of time while computing the relevancy of the web pages.Qiang Zhu et al.[29] proposed OFC on the basis of reinforcement learning and fuzzy clustering theory for the focused web crawler. Naive Bayes classifier is combine with the fuzzy center-averaged clustering method to calculate the fuzzy memberships, used to solve the valLie function mapping the hyperlinks. Online estimation and classification of the newly crawled web pages incrementally enhances the crawling performance.

E. Priority Based Approach

The goal of this paper [25] is to examine the algorithmic aspects of topical crawlers. In this a group of crawling algorithms within our evaluation framework. On the basis of evaluation of crawlers in a specific task it is categorized into two classes of crawling algorithms that are designed and implemented for acting as the best performing crawlers. The main focus of these classes has two typical machine learning issues (i) the role of exploration versus exploitation, and (ii) the role of adaptation (learning and evolutionary algorithms) versus static approaches. The advantage of this approach is to interpret pages and select the links to be visited with the help of decentralization of the crawling process. The author et al.[26] proposes a baseline crawler which is based on a focused crawling approach designed by Soumen Chakrabarti, Martin van den Berg, and Byron Dom. It employs canonical topic taxonomy to train a naïve-Bayesian classifier which help to determine the relevancy of crawled pages. Which URLs is to visit next is based on the locality of topical crawler. A rule-based crawler is used to define simple rules to decide crawler next move which is based on interclass patterns. The rule-based crawler supports tunneling to improve the baseline crawler's harvest rate and coverage. The problem in baseline focused crawler was that it had low relevance scores due to which it might miss future on topic pages, thus as the traversed path is increased the exploration of such path is decreased. This approach increased the downloaded URL by computing the harvest ratio. It continuously retrieves relevant pages better than the previous crawler. M. Ehrig et al.[24] proposes an approach in which building a framework for document discovery on the basis of web documents of ontology focused crawler. The term 'ontology' determines knowledge as a set of concepts within a specific domain, and the relationships among them. It can be used to reason about the entities within that domain and may be used to describe the domain Complex ontology and associated instance elements are used for this framework which defines several relevance computation strategies. These strategies show promising result of an empirical evaluation. The previous crawlers took lot of time in eliminating the irrelevant pages rather than focusing on crawling the relevant pages. The advantage of this approach is that it has high harvest ratio to crawl the relevant pages

efficiently. In the paper [1] author proposed a novel, and distinctive focused crawler named LSCrawler, that retrieves documents based on the keywords in the link and the surrounding text of the link by speculating the relevancy of the document, Which is reckoned by measuring the semantic similarity between the keywords in the link and the taxonomy hierarchy of the specific domain. LS crawler provide better recall as it exploits the semantic of the keywords in the link. The advantage of this approach is to enhance the process of determining the relevancy of the documents before downloading.

F. Miscellaneous Approach

Anuradha et al.[3] proposed Anuradha et al.[3] proposed a novel approach which is used to combine wide range of Web information, that contains dynamically generated Web pages and cannot be indexed by automated Web crawlers that are already being exist, through the knowledge, which is explore from web sources the ontologies can built. Here, Ontology based search is divided into distinct modules. The first constructs attribute-value ontology, second one constructs the attribute-attribute ontology and on the other hand the third module formulate the interface using domain ontology, fills the search, user query, extract results by looking into the index database. Hati et al.[4] proposed an approach, which calculate the unvisited URL score based on its Anchor text relevancy, its description in Google search engine after that calculate the score based on similar description with topic keywords, cohesive text similarity with topic keywords and Relevancy score of its parent pages and vector space model is used for calculating relevancy score. UDBFC Approach:Hati et al.[6] gives UDBFC (URL Distance BasedFocused Crawler) algorithm which is based on a double crawler framework (an experimental crawler and a focusedcrawler) and it is used to calculate the relevancy by using vector space model between seed page and child page. Link extractor tool is used to extract the child page links that are out links of the seed page and experimental crawler is used to fetch seed page and child page. It calculates the relevancy between seed page and its all child pages. Relevancy score defines group on the basis of relevance calculation. It uses the focused crawler to fetch topic specific pages from internet based on distance score which is calculated between grouped URLs and each URL which is to be fetched.

IV. CONCLUSION

Crawlers have always struggled to keep up with Web content generation and modification. A focused crawler or topical crawler is a web crawler that attempts to download only web pages that are relevant to a pre-defined topic or set of topics. They attempt to download pages that are similar to each other. This paper gives a detail of various approaches given by various authors in the past few years. It gives the stage wise development in the field of focused crawling their weaknesses and strengths. So it can be used as a base paper for developing new approaches considering the limitations of existing ones.

REFERENCES

[1] M. Yuvarani; N.Ch.S.N. Iyengar; A. Kannn "LSCrawler: A Framework for an Enhanced FocusedWeb Crawler Based on Link Semantics" Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on Topic(s): Communication, Networking Broadcasting ;Computing & Processing (Hardware/Software).

[2] Yixue Sun; Peiquan Jin; Lihua Yue "A Framework of a Hybrid Focused Web Crawler", Second International Conference on volume 2 Communication, Networking & Broadcasting ,2008,PP-50-53.

[3] Anuradha; Sharma, A.K. "Accessing the Deep Web Using Ontology" 3rd International Conference on Communication, Networking & Broadcasting ;Computing & Processing (Hardware/Software),2010,PP-565-568.

[4] Hati, D.; Sahoo, B.; Kumar, A. "Adaptive focused crawling based on link analysis" 2nd International Conference on volume 4 Communication, Networking & Broadcasting ;Computing & Processing (Hardware/Software) ; Robotics & Control Systems,2010,PP-455-460.

[5] Hati, D.; Kumar, A. "Improved focused crawling approach for retrieving relevant pages based on block partitioning", 2nd International Conference on volume3 Communication, Networking & Broadcasting ;Computing & Processing (Hardware/Software) ; Robotics & Control Systems,2010,PP-269.

[6] Hati, D.; Kumar, A." UDBFC: An effective focused crawling approach based on URL Distance calculation", 3rd IEEE International Conference on volume 3 Bioengineering ; Communication, Networking & Broadcasting ; Computing & Processing (Hardware/Software),2010,PP-59-63.

[7] Jalilian, O.; Khotanlou, H., "A new fuzzy-based method to weigh the related concepts in semantic focused web crawlers" 3rd International Conference on" Volume: 3 Communication, Networking & Broadcasting ;Computing & Processing (Hardware/Software), 2011 , PP-23 – 27.

[8] GOUSIA TABASSUM.;A.POONGODAI.," An Ontology Based Search for Relevant pages using Semantic Web, Search Engines ", (ijaest) International Journal of Advanced Engineering Sciences and Technologies ,Volume11, issue no. 1, pp-106 – 110.

[9] F. Menczer, G. Pant, P. Srinivasan, Topical web crawlers: evaluating adaptive algorithms, ACM Transactions on Internet Technology (TOIT) 4 (4) (2004)378–419.

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

[11] M. Ehrig, A. Maedche, Ontology-focused crawling of web documents, in: Proceedings of the Symposium on Applied Computing (SAC 2003), March9–12, 2003.

[12] A. Hliaoutakis, G. Varelas, E. Voutsakis, E.G.M. Petrakis, E. Milios, Information retrieval by semantic similarity, International Journal on Semantic Web and Information Systems (IJSWIS) 3 (3) (2006) 55–73 (Special issue of multimedia semantics).

[13] G. Pant, P. Srinivasan, Learning to crawl: comparing classification schemes, ACM Transactions on Information Systems (TOIS) 23 (4) (2005) 430–462.

[14] Jun, Li, K. Furuse, K. Yamaguchi, Focused crawling by exploiting anchor text using decision tree, in: Proceedings of the 14th International World Wide Web Conference, 2005, pp. 1190–1191.

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

[16] 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.

[17] Y. Chen, A Novel Hybrid Focused Crawling Algorithm to Build Domain-Specific Collections, Ph.D. Thesis, Virginia Polytechnic Institute and State University, 2007

[18] S. Bri, L. Page, "The anatomy of large-scale hypertext Web search engine", Proc of World-Wide Web Conference, Brisbane, Australia, 1998, 107-117. .paper3

[19] Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", Journal of the ACM, 1999, 46(5), 604-632. .paper3

[20] M. Hersovici, A. Heydon, M. Mitzenmacher, D.pelleg, "The Shark search Algorithm-An application: Tailored Web Site Mapping. Proc of World Wide Conference", Brisbane. Australia, 1998, 317-326. .paper3

[21] A. Pal, D. S. Tomar and S.C. Shrivastava. "Effective Focused Crawling Based on Content and Link Structure Analysis", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 2, No. 1, June 2009.(paper4)

[22] Y. Zhang, C. Yin and F. Yuan. "An Application of Improved PageRank in Focused Crawler", Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007IEEE). .(paper4)

[23] Q. Cheng, W. Beizhan and W. Pianpian. "Efficient focused crawling strategy using combination of link structure and content similarity",IEEE 2008. .(paper4)

[24] M. Ehrig, A. Maedche, "Ontology-focused crawling of Web documents", Proceedings of the 2003 ACM Symposium on Applied Computing, pp. 1174-1178, USA, 2003.(last papr)

[25] F. Menczer, G. Pant, and P. Srinivasan, "Topical web crawlers: Evaluating adaptive algorithms", ACM Transactions on Internet Technologies, Vol.4, No.4, 2004.(last papr)

[26] Ismail Sengor Altingovde and Ozgur Ulusoy, “Exploiting Interclass Rules for Focused Crawling”, IEEE Intelligent Systems, November / December, 2004, pp 66-73(lscrawlr)

[27] M. Jamali, H. Sayyadi, B. B. Hariri, and H. Abolhassani. A method for focused crawling using combination of link structure and content similarity. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, WI ’06, pages 753–756, Washington, DC, USA, 2006. IEEE Computer Society.[5]

[28] Y. Sun, P. Jin, and L. Yue. A framework of a hybrid focused web crawler. Future Generation Communication and Networking Symposia, 2008. FGCNS ’08. Second International Conference on, 2, 2008.[5]

[29] Qiang Zhu “An Algorithm OFC for the Focused Web Crawler.Machine Learning and Cybernetics, 2007 International Conference Vol. 7 Page(s): 4059 –4063.