ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

A Study of Focused Crawler Approaches

Ashwani Kumar, Dr. Anuj Kumar

HOD, Dept. of CS, DIET, Noopur, Bijnor, UP, India

HOD, Dept. of Maths, A.I.M.T, Greater Noida, India

ABSTRACT: A focused crawler is web crawler that focused Crawler main aim is to selectively seek out pages that are relevant to pre-define set of topic rather than to exploit all regions of web. In this paper we present a comparison between simple crawler and focused crawlers as well as various approaches of focused crawling like ontology based, structured based, context based priority based and learning based crawling. In priority based crawling, Priority base crawler assign priority values to URL’s which have been crawled And in structure based crawling, Structured base crawler uses web pages structure to calculate the page relevance. In context based focus crawling, Context based focus crawler also considers context related with topic keyword. In learning based focus crawling, Learning based focus crawler uses classifier to determine whether page is relevant or not.

KEYWORD: Focused crawler, ontology based focus crawler, structure based focus crawler, context based focus crawler, learning based focus crawler, priority based focus crawler.

I.INTRODUCTION

A focused crawler is a computer program that browses the World Wide Web in a methodical, automated manner. They are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. It usually starts with a set of start Uniform Resource Locator (URL) downloads the web page at that location, and recursively retrieves all the pages pointed to by the hyperlinks on the page until it comes to a dead end or until some restriction defined in the crawling policy is met [6] [3]. Increasing volume of the World Wide Web raises the challenges for the search engines and web crawlers [1]. Finding relevant information using a normal search engine retrieves thousands of matches. Going through this size of information is too large; this leads to the problem of information overkill [4]. Using a domain specific search engine to find the relative information will increase the precision to the retrieved information. As a result for using such search engine, a need to design a specific crawler is appeared. This crawler needs to crawl the relevant web pages with very high precision [4]. This domain specific crawling is called focused crawling or topical crawler. Focused crawlers are becoming increasingly important as the growing of the web size [5]. A focused crawler is a Web crawler that attempts to download only web pages that are relevant to a pre-defined topic or set of topics or a specific domain [7]. Focused crawlers vary from general crawlers in that they judge whether the documents pointed to by the URLs are relevant for the specific domain. Ordering the URL queue is based on the relevance probabilities, and the pages assessed to be very relevant for the specific domain are downloaded first [6]. Focused crawlers aim at selectively look for pages that are the relevant web pages and ignore download the irrelevant ones [2] [8]. The benefit of the focused crawling approach is that it is able to find a large number of relevant documents on that specific domain and is able to effectively discard irrelevant documents and hence leading to significant savings in both computation and communication resources, and high quality retrieval results [2]. A focused crawler is web crawler that efficiently gathers Web pages that fulfills specific criteria, by carefully prioritizing the crawl frontiers and managing the hyperlink exploration process. Crawl frontier is the link on a web page that a web crawler can select while performing crawling process. Some predicates may be based on simple or surface properties. A selective crawler's task may be to crawl pages from only the .in domain whereas some other crawler’s aim would be to crawl pages only from .jp domain. While other predicates may be a softer comparative, e.g., crawl pages with large Page Rank or Page Hit, or crawl pages that are related to volleyball.

DOI: 10.15680/IJIRCCE.2015. 0312140

II.COMPARISON

Table 1 Difference between a simple crawler and focused crawler

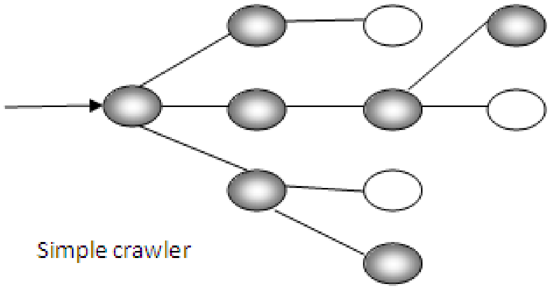

Fig 1

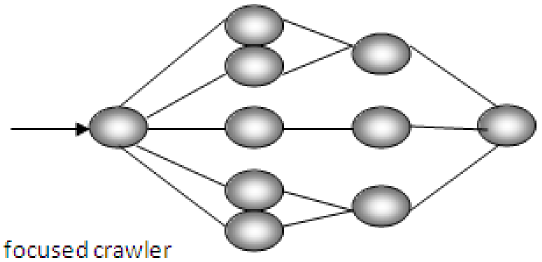

Fig 2

Crawler	**Simple crawler**	**Focused crawler**
Discovered by	Different contributors and sources	S. Chakrabarti, M. v. d. Berg, and B. Dom,
Definition	It locates information on WWW, indexes all the words in a documents and follow each and every hyperlink	It indexes information based on Domains, application and the query that inserted.
Web pages	May or may not be related to each other	Has to be related to a certain domain.
Relevance	Less relevant web pages are discovered.	More relevant web pages are discovered
Consumption of Resources	Consumption of Resources is Less	Consumption of Resources is High
Performance dependency	Independent	Dependent on link richness with in a specific domain.
performance	performance is low	performance is high

III.DIFFERENT APPROACHES USED BY FOCUSED CRAWLER

Focus crawler approaches are usually categorized with respect to their dependency on determining the relevant web pages they focus on-(1) Ontology based focus crawler (2) Structure based focus crawler (3) Context based focus crawler (4) Priority based focus crawler (5) Learning based focus crawler

A. *ONTOLOGY BASED FOCUSED CRAWLER*

Ontology based focus crawlers depend on ontology to determine the relevant page. These are a series of crawlers which use ontologies to link the fetched web documents with the domain concepts [17]. Ontology base focused crawlers relay on utilizing the domain concepts to evaluate the page relevance. Most of researches use ontology to evaluate the relevance before downloading the web pages. Pahal and his colleges [16] and Su and his colleges [13] determine the relevance score for the new download pages according to the relevance of the web contain of the page which has linked to their URL. While Kumar and Vig [2] use the words near to the links to calculate the relevance of the linked URL. Others use ontology to filter the retrieved pages, like [15]. Su and his colleges [13] present an intelligent focused crawler in which embedding ontology is used to evaluate the pages relevance to a topic. Crawler exploits the Web's hyperlink structure to retrieve pages by traversing links from previously retrieve ones. As pages are fetched, their outward links may be added to a list of unvisited pages, which is referring to as the crawl frontier. To identify the next most appropriate link to follow from the frontier, they used an ontology based algorithm to compute page relevance. The ontology concepts are extracted from the download page and are used to calculate the relevance of the page. Then a candidate list of Web pages in order of increasing priority is maintained. Based on the page ranking a decision will be taken to add its links to be crawling or no. They have an ontology learning module in their approached use the download pages to bootstrap their ontology. Pahal and his colleges [16] try to improve the existing focused crawling by using the concept with its context and context information for retrieving web documents.

B.STRUCTURE BASED FOCUSED CRAWLER

Structure base focused crawlers take in accounting the web pages structure when evaluating the page relevance. Jamali and his colleges [8] and Huang and his colleges [14] analysis the hyperlinks between the candidate crawled page and the domain web pages and to determine if it relevant to the domain or not.
Structure base focused crawlers are following-

(1**)** ***Division Score and Link Score based focused crawler***
(*2) Combination of Content and Link Similarity based Focused Crawling*

Division Score and Link Score based focused crawler Debashis Hati et.al [20] proposed an approach in which crawler fetch those link first whose link score is high. However, link score is calculated on the basis of division score and average relevancy score of parent pages of particular link. Here, division score is taken for calculating link score because detailed description of link is available in division in which the link belong. Division score means how many topic keywords belong to division in which the particular link belongs. If all the topic keywords are available in division in which the URL belongs then division score of URL is 1, otherwise it depends upon the percentage value of topic keyword appearance in division. Average relevancy score of parent page is taken for calculating the link score due to following reasons:

* A link from parent page to child page is a recommendation of child page by the author of parent page.
* If parent page and child page are connected by a link the probability that they are on the same topic is higher. Than if they are not connected.
The link_score whose value is greater than threshold is store in queue and link whose link_score is greatest from all is crawled next.

Combination of Content and Link Similarity based Focused Crawling: Jamali et.al [21] uses combination of the link structure analysis and content similarity in building their focused crawling. Their idea is based on that, the ordinary hyperlinks in pages are a representation to the authors view about other pages. Also the contents of pages are another source to relate them to a domain. HAWK: A Focused Crawler with Content and Link Analysis [22] which combines search strategy based on content and link structure. Here Link analysis is based on anchor score, parent score etc.

Context based focus crawler

Context based focus crawler also considers context related with topic keyword. Gupta and his colleges [4] present a framework of a context based distributed focused crawler which maintains an index of web documents relating to the context of keywords resulting in storage of more related documents. Their crawl starts by a list of seed URLs. This list is distributed in multiple crawlers to download the pages corresponding to the. These downloaded pages are indexed by extracting their keywords. Then the system extracts the different contextual interpretations/senses of these keywords from the Word Net dictionary to prepare an index of the local database on the basis of extracted different contextual meaning and senses. Sushil Kumar et.al [23] proposed a context model for focused web search. The previous approach of information retrieval is like a black box; Search system has limited information of user needs. The user context and their environment are ignored resulting in irrelevant search result. This type of system increase overhead to the user in filtering useful information. In fact, contextual relevance of document should also be considered while searching of document.

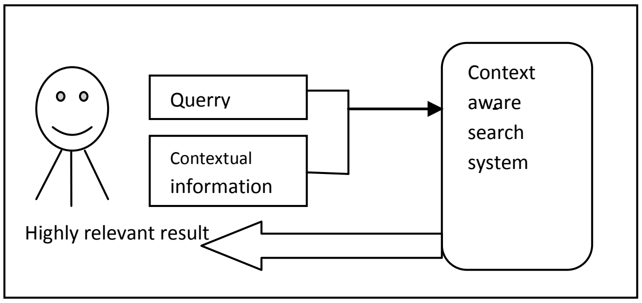

Fig 3 represents the outer-view of contextual driven search system.

The Context aware web Search System has following components:
Capturing the context: The purpose of this component to capture the contextual information from user and environment. Contextual information can be provided by two ways: explicitly by user; here user may be asked to answer some queries when he gives his search query and implicitly by judging user behavior while he interacts with the system, by their profile and by understanding their area of working and environment.
Tagging context: Context of web page must be embedded in document so that while crawling crawler only compare the context of topic with the context of web page. The document index must embed the context so that searching from index, relevant document presented to user. Context derived either explicitly or implicitly from user is augmented with query and submitted to search system. Thus, user gets relevant result according to their query.
Adaption of context: The web search system after capturing and learning the contextual information about the user and his environment must adapt it to the users actual needs presents highly relevant search results to him**.**

Priority Based Focused Crawler

Jaytrilok choudhary et.al [18] proposed priority based focused crawling. This crawler assign priority values to URL's which have been crawled. This type of crawler keeps the relative score along with the corresponding URL's to be visited in a priority queue. When a URL from a priority queue is deleted, it returns a maximum score URL. The web page corresponding to URL is downloaded from web and calculates the relative score of download page with focus word. Here, URL extracted from a page is stored in priority queue instead of normal queue. Thus, every time crawler return the maximum score URL to crawl next. This type shows about 85% improved results over simple crawler .One drawback of this method is that it is very time consuming and can be eliminated by implementing parallel algorithm.

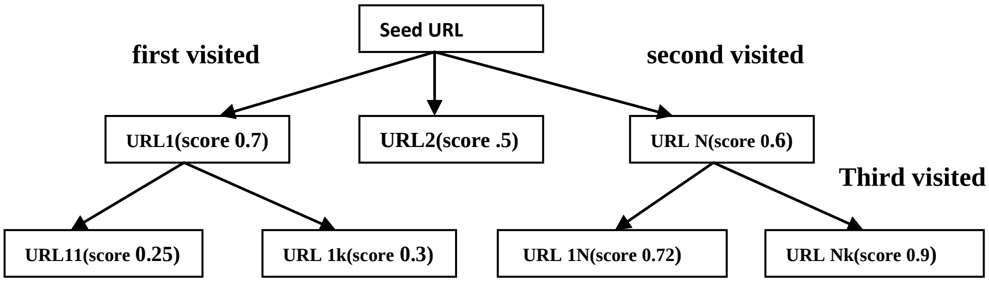

Fig.4 crawler start with seed URL and URL1 whose score is 0.7 is visited first after that URL N, URL NK and so on.

Learning based focus crawler

Learning based focus crawler uses classifier to determine whether page is relevant or not. S.Safran et.al [19] proposed a new learning based approach to improve relevance prediction in focused web crawler. Firstly, training set is built to train the system .Training set contain value of four relevance attributes: URL word relevancy, anchor text relevancy, parent page relevancy, and surrounding text relevancy. Secondly they train the classifier (NB) using training set. After that trained classifier is used to predict the relevancy of unvisited URL.

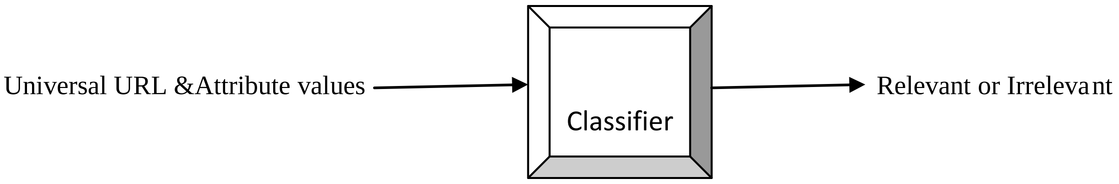

Fig. 5 Classifier input and output

Training set preparation: To train the classifier both relevant and irrelevant URLs are needed. After that attributes relevance value is calculated like URL word relevancy, anchor text relevancy, parent page relevancy etc. of both relevant and irrelevant URLs.

Training and Prediction: NB classifier is trained on the training set and used to predict the relevancy of unvisited URL. The method that is used to calculate the probability of given URL as relevant or irrelevant is given below**:**

$P(N|R=y)*P(R=y) > P(N|R=n)*P(R=n)$ (1)

Where N is unvisited URL .The left side of equation considered N is relevant and right side considered N is irrelevant. If given equation is true then N is relevant, otherwise irrelevant**.** The way of computing term is explained below in eq. (2), eq. (3), and eq. (4) $P(N|R=y) = P(URL\_word\ relevancy \mid R= y) * P(anchor\_text\ relevancy \mid R=y) * P(parent\_page\ relevancy \mid R=y)*P(surround\_text\ relevancy \mid R=y)$ (2)

$P(R=y)$ = Number of relevant pages/Total pages (3)

$P(R=n)$ = Number of irrelevant pages/Total Pages (4)

Where y=yes=no and r= relevant

IV.CONCLUSION

Web crawling is one of the main component in applications like search engines. We compared between standard and focused web crawlers to understand which one is better and also discussed about the merits of various approaches like Ontology based ,Structure based, Learning based and priority based as well as contextual based focused crawling. The advantages of focused crawler are that we spend less money, time & effort processing web pages that are most unlikely to be of value or worth.

REFERENCES

[1] A. Thukral, V. Mendiratta, A. Behl, H. Banati and P. "Bedi, FCHC: A Social Semantic Focused Crawler", in Communications in Computer and Information Science, Vol. 191, Part 5, pp. 273-283, 2011.

[2] M. Kumar and R. Vig, "Design of CORE: context ontology rule enhanced focused web crawler", International Conference on Advances in Computing, Communication and Control (ICAC3‟09) pp. 494-497, 2009.

[3] A. Chandramouli, S. Gauch, and J. Eno, "A Cooperative Approach to Web Crawler URL Ordering", iIn Human Computer Systems Interaction, AISC 98, Part I, pp. 343–357, 2012

[4] P. Gupta, A. Sharma, J. P. Gupta, and K. Bhatia, „A Novel Framework for Context Based Distributed Focused Crawler (CBDFC)‟, Int. J.CCT, Vol. 1 , No. 1 , pp.13-26. 2009

[5] A. Patel, and N. Schmid, "Application of structured document parsing to focused web crawling", in Computer Standards & Interfaces 33 (2011) pp. 325–331

[6] A. Pirkola and T. Talvensaari, “Effects of Start URLs in Focused Web Crawling", in INFORUM 2009: 15th Conference on Professional Information Resources Prague, May 27-29, 2009.

[7] S. Yang and C. Hsu, “An Ontology-Supported Web Focused-Crawler for Java Programs”, Proc. of 2010 International Workshop on Mobile Systems, E-commerce, and Agent Technology, Jinhua, China, Jul. 5-6, 2010, pp. 266-271

[8] M. Jamali , H. Sayyadi , B. B. Hariri, and H. Abolhassani, “A Method for Focused Crawling Using Combination of Link Structure and Content Similarity”, Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pp.753-756, December 18-22, 2006.

[9] S. Chakrabarti, M. v. d. Berg, and B. Domc, “Focused crawling: a new approach to topic-specific Web resource discovery”, Computer Networks, 31(11–16):1623–1640. 1999

[10] A. Pirkola, “Focused Crawling: A Means to Acquire Biological Data from the Web”, in VLDB ‟07, September 23-28, 2007, Vienna, Austria.

[11] A. Micarelli and F. Gasparetti, “Adaptive Focused Crawling”, in The Adaptive Web, LNCS 4321, pp. 231–262, 2007.

[12] Q. Xu and W. Zuo, “First-order Focused Crawling”, WWW 2007, pp. 1159-1160.

[13] C. Su, Y. Gao, J. Yang, and B. Luo, “An efficient adaptive focused crawler based on ontology learning”, Hybrid Intelligent Systems, 2005. HIS apos;05. Fifth International Conference on 6-9 Nov. 2005.

[14] W. Huang, L. Zhang, J. Zhang, M. Zhu, "Focused Crawling for Retrieving E-commerce Information Based on Learnable Ontology and Link Prediction" ieec, International Symposium on Information Engineering and Electronic Commerce, pp.574-579, 2009.

[15] H. P. Luong, S. Gauch, and Q. Wang, “Ontology-Based Focused Crawling”, Information, Process, and Knowledge Management, 2009 (eKNOW '09) 1-7 Feb. 2009 pp. 123-128

[16] N. Pahal, N. Chauhan, and A.K. Sharma, “Context-Ontology Driven Focused Crawling of Web Documents”, A.K. Wireless Communication and Sensor Networks, 2007. WCSN apos:07. Third International Conference, 13-15 Dec. 2007 pp.121-124

[17] H. Dong, F. K. Hussain, and E. Chang, "State of the art in semantic focused crawlers" in 2009 IEEE International Conference on Industrial Technology (ICIT 2009), Gippsland, in press

[18] Jaytrilok Choudhary and Devshri Roy ,“ A Priority Based Focused Web Crawler” , International Journal of Computer Engineering and Technology , Volume 4 ,Issue 4, july-august 2013.

[19] Mejdl S. Safran, Abdullah Althagafi and Dunren Che , “Improving Relevance Prediction for Focused Web Crawlers” , IEEE/ACIS 11th International Conference on Computer and information Science.

[20] Debashis Hati , Amritesh Kumar ,” An Approach for Identifying URLs Based on Division Score and Link Score in Focused Crawler” ,International Journal of Computer Applications , Volume 2 – No.3, May 2010.

[21] Brin, S. and Page, L. (1998),“The Anatomy of a Large- Scale Hypertextual Web Search Engine,” Computer Networks and ISDN Systems, 30(1–7).

[22] X.Chen and X. Zhang , “HAWK: A Focused Crawler with Content and Link Analysis”, Proc. IEEE International Conf. on e-Business Engineering ,2008.

[23] Sushil Kumar ,Naresh Chauhan , “A Context Model For Focused Web Search”, International Journal of Computers & Technology Volume 2 No. 3, June, 2012.