Volume 4, Issue 7, July 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review of Focused Crawler Approaches

Meenu
Dept. of Computer Science
YMCA, Faridabad, India

Rakesh Batra
Dept. of Computer Science
YMCA, Faridabad, India

Abstract - 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 a review of focused crawler approaches have been presented which is classify in to five categories: Priority base crawler, Structured base crawler, Leaning base crawler, Context base crawler and Other focused crawler. Priority base crawler assign priority values to URL's which have been crawled. Structured base crawler uses web pages structure to calculate the page relevance. Learning base crawler uses classifier to determine whether page is relevant or not. Context base crawler also considers context related with topic keyword. Other Focused Crawler can not be classified to one of the previously focused crawler approaches.

Keywords-Crawler; Focused Crawler; Structured base crawler; Learning base crawler; Context base crawler

I. INTRODUCTION

World Wide Web contains a large amount of information. And in every second, new information is added. Thus to find relevant information on WWW is very difficult task. Search Engine overcomes this problem. It automatically visits web sites and create index to enable searching for information. Web Crawler is the main component of search engine. It continuously downloads pages from WWW. These pages are indexed and stored in database. Recent study estimates the size of web that passed over billions of documents. It becomes impossible for a crawler to crawl whole web and keep its index fresh. Thus there is need of crawler which crawl only relevant subset of WWW. This crawler is known as focused crawler. Focused Web crawlers traverse the Web by exploiting its link structure like generic web crawlers. Focused crawler select relevant pages from WWW that are related to predefine topic and ignore the downloading of irrelevant page. Focused crawler is also known as domain specific crawler.

In this chapter a survey of different approaches of focused crawling has been described. The outline of this paper is as follows: section 2 describes a brief description of focused crawler. Section 3 describes various focused crawling approaches and in section 4 conclusions have been presented.

II. FOCUSED CRAWLER

The term focused crawler was first introduced by Chakrabarti et.al [1]. They described the focused crawler in which a crawler seeks, acquires, indexes, and maintains pages on a specific set of topics that represent a relatively narrow segment of the web [1].The Fish-Search [12] is an example of early crawlers that prioritizes unvisited URLs on a queue for a specific search goal. The Fish-Search approach assigns priority values (1 or 0) to candidate pages using simple keyword matching. One of the disadvantages of Fish-Search is that all relevant pages are assigned the same priority value 1 based on keyword matching. The Shark-Search [12] is a modified version of Fish-Search, in which, Vector Space Model (VSM) is used, and the priority values (more than just 1 and 0) are computed based on the priority values of parent pages, page content, and anchor text.

Info Spiders and Best-First are additional examples of focused crawling methods [9]. The difference between them is that Info Spiders uses Neural Networks, while Best-First method applies VSM to compute the relevance between candidate pages and the search topic. Best-First was shown most successful due to its simplicity and efficiency. *N*-Best-First is generalized from Best-First, in which *N* best pages are chosen instead of one. There are so many approaches of focused crawling which is explained in next section.

III. DIFFERENT APPROACHES OF FOCUSED CRAWLING

Focused crawler approaches can be categorized according to their dependency on determining the relevant pages to: priority based focused crawler, structure based focused crawler, context based crawler, learning based crawler and others focused crawler approaches.

A. Priority Based Focused Crawler

Jaytrilok choudhary et.al [11] proposed priority based focused crawling .The web page corresponding to URL are downloaded from web and calculate 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.

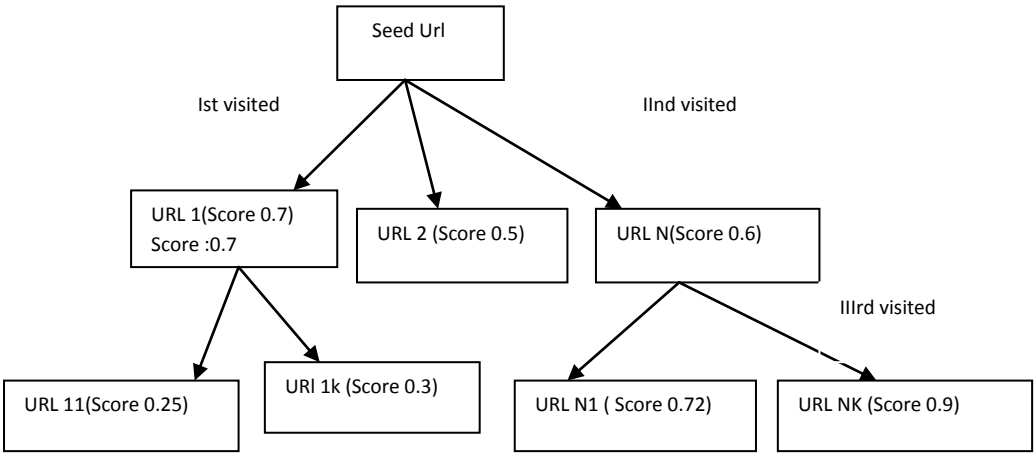

Fig.1 Priority Based Focused Crawling Process

As shown in fig.1 crawler start with seed URL and URL1 whose score is 0.7 is visited first after that URL N, URL NK…. and so on.

B. Structure Based Focused Crawler

Structure base focused crawlers take in accounting the web pages structure when evaluating the page relevance. Some structure based focused crawler are described below:

1) *Division Score and Link Score based focused crawler:* Debashis Hati et.al [10] 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.

2) Combination of Content and Link Similarity based Focused Crawling: Jamali et.al [8] 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 [5] which combines search strategy based on content and link structure. Here Link analysis is based on anchor score, parent score etc.

C. Context Based Focused Crawling

Sushil Kumar et.al [7] 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 2 represents the outer-view of contextual driven search system.

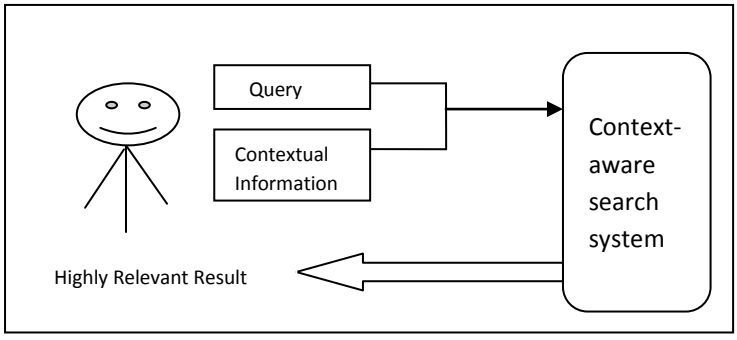

Fig. 2 Context driven search system

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**.**

D. Learning Based Crawler

S.Safran et.al [3] 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.

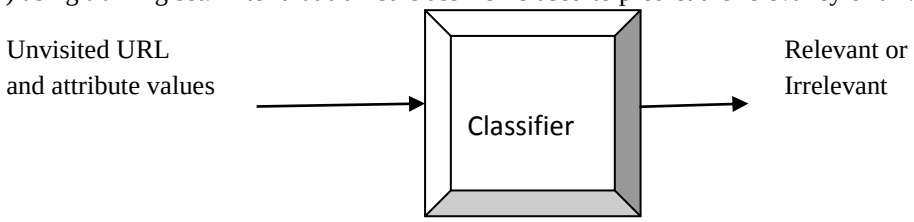

Fig. 3 Classifier input and output

- *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|Relevant=yes)*P(Relevant=yes) > P(N|Relevant=no)*P(Relevant=no) (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|Relevant=yes) = P(URL_word relevancy | Relevant= yes) * P(anchor_text relevancy | Relevant=yes) * P(parent_page relevancy | Relevant=yes)*P(surrounding_text relevancy | Relevant=yes) (2)

P(Relevant=yes) = Number of relevant pages/Total pages (3)

P(Relevant=no) = Number of irrelevant pages/Total Pages (4)

E. Other Focused Crawler

In this section, those focused crawlers are presented which have their own features that cannot be categorized to one of the previously focused crawlers approached.

Bazarganigilan et.al [6] present a focused crawler that use similarity function to determine the page relevancy. They use genetic programming to discover the best combination for estimation the similarity evaluation among pages. Their crawler download the web pages pointed to by the starting URLs. For each downloaded web page, the similarity function will be used by a classifier to determine if this is a computing-related Web page. If yes, this web page will save into the download collection. The outgoing links of the download relevant web pages will be collected and insert into the crawling queue. They apply a decay concept to each page. The page would stop of crawling if it does not comply with predefined threshold.

Gupta et.al [4] present a framework of a context based distributed focused crawler. Their crawl starts by a list of seed URLs. This list is distributed in multiple crawlers to download the pages. 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.

IV. CONCLUSION

General Crawler has some limitation in terms of precision and efficiency because of its generality, no specialty. Focused Crawler improves the precision and recall of expert search on web. Focused crawler does not collect all pages but select and retrieve relevant page only. There are so many approaches to calculate the relevancy of page. Some base on structured, some used classifier to know the relevancy of page etc. Context based focused crawling give more accurate result to user according to their interest.

REFRENCES

[1] S. Chakrabarti, M. Van Den Berg, B. Dom, "Focused Crawling: A New Approach to Topic specific web resource discovery", Proc. Of 8th International WWW conference, Toronto, Canada, May,1999.

[2] Ayoub Mohamed H. Elyasir1, Kalaiarasi Sonai Muthu Anbanan, "Focused Web Crawler", International Conference on Information and Knowledge Management, 2012.

[3] 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.

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

[6] M. Bazarganigilani, A. Syed and S. Burki, "Focused web crawling using decay concept and genetic programming", In International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.1, pp:1-12, 2011.

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

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

[9] X. Zhang, T. Zhou, Z.Yu and D.Chen, "URL Rule Based Focused Crawlers", IEEE International Conference on e-Business Engineering, 2008.

[10] 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.

[11] 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.

[12] Novak, B., "A survey of focused web crawling algorithms", in Proceedings of SIKDD 2004 at Multiconference IS. 2004, ACM Press: Slovenia. p. 55-58.

