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. 4, Issue. 5, May 2015, pg.545 – 551

RESEARCH ARTICLE

Efficient Focused Web Crawling Approach for Search Engine

Ayar Pranav[1], Sandip Chauhan[2]

Computer & Science Engineering, Kalol Institute of Technology and Research Canter, Kalol, Gujarat, India

[1] pranavayar@gmail.com; [2] sandymba2006@gmail.com

Abstract— a focused crawler traverses the web, selecting out relevant pages to a predefined topic and neglecting those out of concern. Collecting domain specific documents using focused crawlers has been considered one of most important strategies to find relevant information. While surfing the internet, it is difficult to deal with irrelevant pages and to predict which links lead to quality pages. However most focused crawler use local search algorithm to traverse the web space, but they could easily trapped within limited a sub graph of the web that surrounds the starting URLs also there is problem related to relevant pages that are miss when no links from the starting URLs. There is some relevant pages are miss. To address this problem we design a focused crawler where calculating the frequency of the topic keyword also calculate the synonyms and sub synonyms of the keyword. The weight table is constructed according to the user query. To check the similarity of web pages with respect to topic keywords and priority of extracted link is calculated.

Keywords— Web crawler; Focused Crawler; Architecture of focused web crawler; Different approaches of focused web crawler; various focused web crawling stratagems; existing method; proposed method, experiment results, conclusion.

I. Introduction

A Web crawler [5] is a key component inside a search engine [11]. Web crawling is the process by which we gather pages from the Web, in order to index them and support a search engine. The objective of crawling is to quickly and efficiently gather as many useful web pages as possible, together with the link structure that interconnects them. Web crawlers 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. Because of limited computing resources and limited time, focused crawler has been developed. A focused crawler is web crawler that attempts to download only web pages that are relevant to pre defined topic or set of topic. A focused crawler is called the topical crawler because fetch only those pages that are topic specific. A focused crawler tries to get the most promising links, and ignore the off- topic document.

II. Architecture of focused web crawler

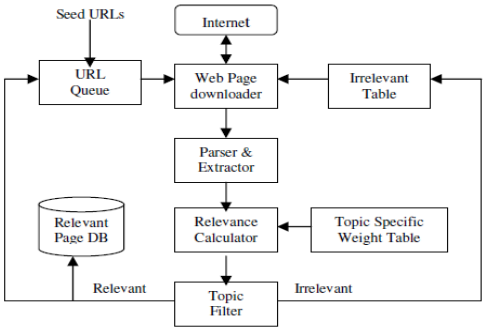

Figure 1 Architecture of focused web crawler [1]

The Architecture of the focused Web crawling is explained here, in the architecture URL Queue contains the seed URLs maintained by the crawler and is initialized with unvisited URLs. Web page Downloader fetches URLs from the URL Queue and Downloads corresponding pages from the internet. The parser and extractor extract information such as the text and the hyperlink URLs from a Downloaded page. Relevance calculator calculates relevance of a page with respect to topic and
assigns score to URLs extracted from the page. Topic filter analyzes whether the content of parsed pages is related to topic or not. If the page is relevant, the URLs extracted from it will be added to the URL queue, otherwise added to the irrelevant table.

III. Different approaches of Focused Web crawling

- **Priority Based Focused Crawler**

The web page corresponding to URL is downloaded from the web and calculates the relative score of download page with focus word. Here, URL get from a page is stored in the priority queue instead of normal queue. Thus every time crawler return the maximum score URL to crawl next.

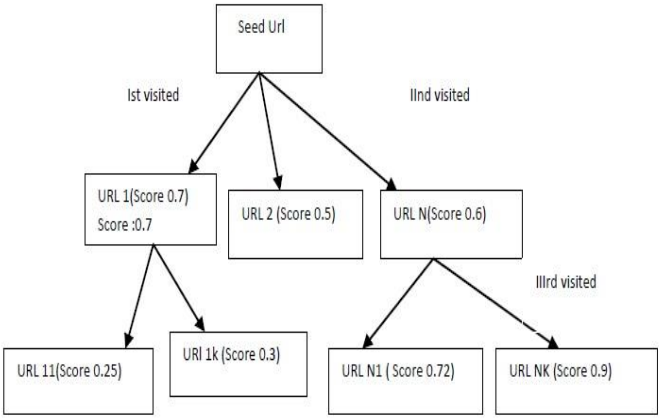

Figure 2 Priority Based Focused Crawling Process [6]

- **Structure Based Focused Crawler**

In the structure base focused crawler the web page structure is taken in accounting when evaluating the page relevance. Some structure based focused crawler are explained below:

- **Division Score and Link Score based focused crawler**

Crawler fetch only 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.

- **Combination of Content and Link Similarity based Focused Crawling**

The Content- Based method is using the page texture information to determine whether the page is pertinent to the topic, and to evaluate the value of page. The famous method of this kind of technology likes the fish-search algorithm and the Shark-search algorithm. The Link-Structure-Based method is

analyzing the reference-information among the pages to evaluate the page value. This kind of famous algorithms likes the Page Rank algorithm and the HITS algorithm [7].

3) Context Based Focused Crawling

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.

4) Learning Based 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.

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.

IV. Various method of focused web crawler

1) Breadth-First Crawling [12]: This is the simplest crawling method in this method retrieve all the pages around the starting point before following links further away from the start. This is the most common approach where robot or crawlers follows all links. If the crawler is indexing several hosts, then this approach distribute the load quickly so we implement the parallel processing.

2) Depth-First Crawling [13]: In Depth-first crawling follow all links from the first link on the starting page, and the follow the first link on the second page, and this process continue. Once the first page is indexed than follow the first link of second page and subsequent links, and follows them. Some unsophisticated use this kind of method, as it might be easier to code.

3) Fish Search: The web is crawled by a team of crawlers, which are viewed as a school of fish. If the fish finds a relevant page based on the keywords specified in query, it continues looking by following more links from that page. If the page is not relevant, then his child links receive low preferential value.

4) Shark Search: It is the modification of fish search. It is differing in two ways: a child inherits a discounted value of the score of its parent, and this score is combined with a value based on the anchor text that occurs around the link in the web page.

5) Naïve best First method [14]: It exploits the fact that if the relevant page links to the other relevant page. Therefore the relevance of a page A to topic t, pointed by page B, is estimated by the relevance of page B to the topic t. each page represented as a vector of weights corresponding to the normalized frequencies of the document's terms according to the TF-IDF scheme. In this method the term frequency is calculated which is the frequency of the term within a document and inverse document frequency where in how many document the term occur.

6) Page Rank Algorithm [15]: It determines the importance of the web pages by counting citations or back links to a given page. The page rank is calculates as:

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1)) + \ldots + PR(Tn)/C(Tn)$$

Where, PR (A) = Page Rank of a Web site,

D = Damping factor.

T1…..Tn = links.

7) HITS algorithm [16]: This algorithm put forward by Kleinberg is previous to Page rank algorithms which uses scores to calculate the relevance. This method retrieves a set of results for a search and calculates the authority and hub score within that set of results. Because of these reasons this method is not often used.

8) Info Spider [17]: Info Spiders complement traditional index based search engines using agents at the user side. These agents act autonomously with each other and they try to achieve a good coverage of the relevant documents. When the user submits a query, Info Spiders obtain a set of seed links which are the search results of a traditional search engine. An agent is initialized for every link and analyses the corresponding page's links looking for the next one to follow. The agent analyses the links by computing the similarity of the text around the link with the query, with the help of a neural net. The next link to be followed is chosen with a probability proportional to the similarity score. The neural net weights are adjusted by the relevance of the new page's content so that the agent updates its knowledge.

9) Intelligent crawling [18]: This method involves looking for specific features in a page to rank the candidate links. These features include page content, URL names of referred Web page, and the nature of the parent and sibling pages.

10) Ontology based focused crawling [19]: In the process of crawling they use ontology. It consists of two main processes which interact with each other. The two main processes is ontology cycle and crawling cycle. In the ontology cycle, the crawling target is defined by ontology (provided by the user) and the documents that are considered relevant as well as proposals for the enrichment of the ontology are returned to the user. The crawling cycle retrieves the documents on the web and interacts with the ontology to determine the relevance of the documents and the ranking of the links to be followed.

11) Metadata based focused crawling [20]: The purpose of the crawler was to harvest missing documents of digital library collections. The crawler could therefore be used to build a complete collection of documents of a given venue i.e. a journal or a conference. The document's metadata are used to locate the home pages of the authors, which are then crawled in order to find the target.

12) Language focused crawling [22]: It uses the language classifier which determines whether page is worth preserving, is incorporated into the crawling process. The crawler is build for the creation of topic specific corpora of a give language. This is the two step process in the first step, a training set of documents which satisfy the language and topic requirement is created in order to extract the most distinguishing ngrams. In the second phase, a classifier is incorporated in the crawler.

V. Existing method

Topic Specific Weight Table Construction [1]

The Weight table defines the crawling target. The topic name is sent as a query to the Google web search engine and the first k results are retrieved. The retrieved pages are parsed, stop words such as 'the' and 'is' are eliminated, words are stemmed using the porter stemming algorithm and the term frequency and document frequency of each word is calculated. The term weight is computed as W = TF*DF. In the next step order the word by weight and extract of words with high weight as the topic keywords. After that weight are normalized as W= Wi / Wmax.

Page Relevance [2]

This module calculates the relevancy of page corresponding to topic keyword in the table by using equation (4). Here, it uses cosine similarity to calculate the relevancy of page:

$$\text{Relevancy}(t, p) = \sum Wkt * Wkp \,/\, \sqrt{\sum} Wkt^2 * \sqrt{} Wkp^2 \quad \ldots\ldots\ldots\ldots (4)$$

Where , CWi(t) and CWi(p)are the weight of i-th common keyword in weight table t and web page p respectively, and Wi(t) and Wi(p) are the weight of keyword in web page p and weight table t respectively. If the relevancy score of page is greater than threshold value then Link Score of its extracting links are calculated by using equation (5).

$$\text{Link Score}(k) = \alpha + \beta + \gamma + \infty \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Where LinkScore (k) is score of link k, α= URLScore (k) is the relevancy between topic keywords and href information of k, β = AnchorScore (k) is the relevancy between topic keywords and anchor text of k, γ=ParentScore (k) is the page relevancy score of page from which link was extracted and ω=SurroundingScore (k) is the relevancy between text surrounding the link and topic keyword. The links whose score is greater than threshold is considered to be relevant. Relevant URLs and their score is stored in relevant URL buffer and signal is given to process URL seen test.

VI. Step-wise Proposed Method

To overcome the problem of missing the some relevant pages some modification is done in the content analysis method. Here in proposed method we include the synonyms and sub synonyms of particular term while calculating the term frequency. In the methodology which is describe here is basically the web analysis method. First is the content based analysis where the content of the web page is consider for the relevance calculation and the link base method also count the relevance of the page by considering the links of the relevant web page.

Step 1:

- Scan the database and get
  - The link from the database and download the all web page content
  - Fetch the number of hyper link of web pages.

Step 2:

- Calculate the relevance calculation
  - Calculating the term weight using the term frequency and inverse document frequency method by using the formula.
  - While calculating the term frequency the synonyms and sub synonyms of the keyword is included in the term frequency.

$$Wi = tf * df$$

  - Normalize the weight by the given formula

$$Wi+1 = Wi / Wmax$$

  and constructing the topic weight table construction

Step 3:

- Calculate the relevance calculation
  - Calculate the topic relevancy of page corresponding to topic keyword in the table by using the equation.

$$Relevancy\ (t, p) = \sum Wkt*Wkp / \sqrt{\sum Wkt^2 * Wkp^2}$$

Step 4:

- Link Ranking calculation.
  - The Links Ranking assigns scores to unvisited Links extracted from the downloaded page using the information of pages that have been crawled and the metadata of hyperlink. Metadata is composed of anchor text and HREF information.

$$LinkScore(k)=\alpha + \beta + \gamma + \infty$$

  Where Link Score (k) is score of link k, α= URLScore (k) is the relevancy between topic keywords and href information of k, β = Anchor Score (k) is the relevancy between topic keywords and anchor text of k, γ=Parent Score (k) is the page relevancy score of page from which link was extracted and ω=Surrounding Score (k) is the relevancy between text surrounding the link and topic keyword.

VII. Experiment Results

The experiments are conducted in Java environment. Breadth-First Search (BFS) crawler is also implemented for performance comparison. There is mysqlconnection.jar file is used for database connection. Jsoup is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data. In order to evaluate the performance of algorithm, we use precision to estimate the efficient of a focused crawling strategy. It is the ratio of topic pages in all of downloaded pages. The formula is shown as follows:

Precision rate= relevant pages/ total downloaded page

After applying the propose step on seed URL and comparing results with focused and BFS algorithm we can say that this method gives more precision results. As number of term frequency is increase so the weight of the keyword is increase and the relevancy of the web page is increase so the number of relevant web page is increase.

VIII. Conclusion and Future Work

This paper, presented a method for focused web crawling that allows to the crawler to go through several relevant pages are missing. From the above step explain in the proposed method better performance than existing method.

Although the initial results are encouraging, there is still a lot of work to do for improving the crawling efficiency. A major open issue for future work is to do extension test with large volume of web pages. Future work also includes code optimization and URL queue optimization because crawler efficiency is not only depends to retrieve maximum number of web page. The dependency of the proposed method is accuracy of the dictionary used in the method.

References

[1] Qu Cheng, Wang Beizhan, Wei Pianpian, "Efficient Focused Crawling Strategy Using Combination of Link Structure and Content Similarity", Software School, Xiamen University, Xiamen 361005, Fujian, China, Proceedings of 2008 IEEE International Symposium on IT in Medicine and Education, 978-1-4244-2511-2/08/$25.00 ©2008 IEEE.

[2] Meenu, Priyanka Singla, Rakesh Batra, "Design of a Focused Crawler Based on Dynamic Computation of Topic Specific Weight Table" International Journal of Engineering Research and General Science Volume 2, Issue 4, June-July, 2014 ISSN 2091-2730.

[3] Anshika Pal, Deepak Singh Tomar, 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.

[4] Bireshwar Gangly, Rahila Sheikh, "A Review of Focused Web Crawling Strategies" International Journal of Advanced Computer Research, volume 2, number 4 issue 6, December 2012.

[5] Jaira Dubey, Divakar Singh, "A Survey on Web Crawler", International Journal of Of Electrical, Electronic and Computer System, ISSN (Online): 2347-2820, Volume-1, Issue -1, 2013.

[6] Meenu, Rakesh Batra, "A Review of Focused Crawler Approaches", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 7, July 2014

[7] Deepali Dave, "Relevance Prediction in Focused Crawling: A Survey", Journal of Information, knowledge and research in computer Engineering, ISSN: 0975 – 6760, Volume 2, issue -2, November 2013.

[8] Debashish, Amritesh, Lizashree "Unvisited URL Relevancy Calculation in Focused Crawling based on Naïve Bayesian Classification", International Journal of Computer Application, volume 3, July 2010.

[9] Anshika Pal, Deepak Singh Tomar, S.C. Shrivastava, "Effective Focused Crawling Based on Content and Link Structure Analysis" Available at: http://arxiv.org/ftp/arxiv/papers/0906/0906.5034.pdf

[10] D. Minnie, S.Srinivasan, "Intelligent Search Engine Algorithms on Indexing and Searching of Text Document using Text Representation" available at: "http://ieeeexplore.ieee.org/xpl/login.jsp".

[11] "About Search engine Optimization" available at: "http://static.googleusercontent.com/media/www.google.com/en//webmasters/docs/search -engine-optimization-starter-guide.pdf"

[12] Najork, M. and Wiener, J., L., 2001. Breadth-First Search Crawling Yields High-Quality Pages. In 10th International World Wide Web Conference, pp. 114-118

[13] Alexander Shen "Algorithms and Programming: Problems and solutions" Second edition Springer 2010, Pg 135.

[14] Grigoriadis, A. and Paliouras, G., 2004. Focused Crawling using Temporal Difference-Learning. In Proceedings of the Panhellenic Conference in Artificial Intelligence (SETN), Samos, Greece, pp. 142-153.

[15] Pant, G. and Menczer, F., 2003. Topical Crawling for Business Intelligence. In Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries.

[16] Brin, S. and Page, L., 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In Proceedings of the seventh international conference on World Wide Web 7.Brisbane, Australia pp. 107 - 117.

[17] Page, L., Brin, S., Motwani, R. & Winograd, T., 1998. The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Library Technologies Project.

[18] Kleinberg, M. J., 1997. Authoritative Sources in a Hyperlinked Environment, In Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithm.

[19] Aggarwal, C., Al-Garawi, F. & Yu, P., 2001. Intelligent Crawling on the World Wide Web with Arbitrary Predicates, In Proceedings of the 10th international conference on World Wide Web, Hong Kong, Hong Kong, pp. 96 – 105.

[20] Ehrig, M. and Maedche, A., 2003. Ontology-Focused Crawling of Web Documents. In Proceedings of the Symposium on Applied Computing 2003 (SAC 2003), Melbourne, Florida, USA, pp. 1174-

[21] Zhuang, Z., Wagle, R. & Giles, C. L., 2005. What‟s There and What‟s Not? Focused Crawling for Missing Documents in Digital Libraries. In Joint Conference on Digital Libraries, (JCDL 2005) pp. 301-310.

[22] Madelyn, O., Schulz, S., Paetzold, J., Poprat, M. & Markó, K., 2006. Language Specific and Topic Focused Web Crawling. In Proceedings of the Language Resources Conference LREC 2006, Genoa, Italy.

