International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Filters Based Focus Crawler Using Utility Theory

Manpreet Kaur[1], Yasmeen Kaur Dhaliwal[2]

[1]Student, Department of Computer Science and Engineering, Chandigarh Engineering College, Landran (Mohali), India

[2]Assistant Professor, Department of Computer Science and Engineering, Chandigarh Engineering College, Landran (Mohali), India

Abstract: ***In the research work, we have developed a filters based focus crawling algorithm and a scoring system in which the relevant web pages are retrieved by the crawler. In this proposed algorithm, three filters such as ego, degree and partition filters are used to find the relevant web pages for crawling. These filters are used for comparing crawled webpage tokens with the master graphs, which reflects the interconnection and closeness between the words between considered and web page that is executed by the crawler. This algorithm calculates high precision, recall value as compared to the best first search algorithm.***

Keywords: Ego Filter, Degree Filter, Partition Filter, Multi domain, Best first search

1. Introduction

Due to rapid growth of the World Wide Web, finding the relevant information would be a challenging and time consuming task. In order to solve this problem, Focused crawlers is developed, [1] as the ideal solution, through mining of the web, which helps us to find the pages from the web that are closely relevant to the desired information. For this purpose, varieties of methods are designed and implement. In this research work, a focus crawler is designed in which filters such as ego, degree and partition are used to find the first node which is to be expanding from the graph and to build a sub graph and ratio is calculated. The ratio reflects the proportion of webpage content relevant to the master graph representing the keywords, tokens interconnections of that particular domain like sports. The filters build the sub graph based the master graph that reflects the interconnections and closeness between the words between considered and web page URL to be executed next. This crawler is multi domain specific, in which domains are, used for example sports, politics etc. The algorithm improves the results of recall, precision values as compare to the previous algorithm.

2. Related Work

Focus crawler is firstly introduced by S. Chakrabarti [1] in 1999. The fish search algorithm, is developed by D. bra [2] which is query driven; starting from a set of seed pages and it considers only those pages that have content matching a given query and their neighborhoods. After that the Shark search algorithm [3] is developed. Which is a modification of Fish Search algorithm which differs 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. Ali pesaranghader [4] developed Term Frequency-Information Content (TF-IC) to prioritize terms in a multi-term topic accordingly, through conducted experiments; it compares its measure against both Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Semantic Indexing (LSI) measures applied in focused crawlers. The Best first Search algorithm [5] is described, which retrieved the relevant web pages. But it crawls from the whole web and take lot of time and resources to retrieve relevant page.

3. Implementation

The figure 3.1 shows the overall flow on which work is done.

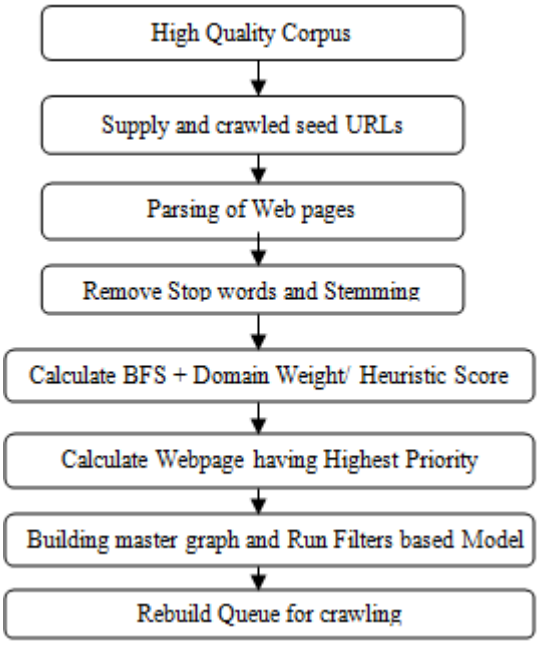


Figure 3.1: Overall flow of work

Step 1: In the first and foremost step in this endeavor is to select high quality corpus. It means that selection of the domain specific keywords. After lot of thinking, the research agenda has been limited to the area that incorporates domain and strategic knowledge. A dictionary of all the key words relating to domain sports and politics is prepared. The proposed master graph has more than 10,000 elements in it specific to particular domain.

Step 2: The crawler is initially supplied with seed URLs which are specific to the domain. The seed URLs are the URLs corresponding to the domain for which the crawler is being built. In this work, the corpus consists of pages related to sports and politics. After supplying the seed URLs, crawling is done**.** The data that is crawled can be news, blog etc. [6]

Step 3: The documents present in the corpus are tokenized in this step. All the files in the corpus are decomposed into tokens. [6]

Step 4: In this step, stop words (such as is, of, being, to, the etc) and the numerical values are removed from the corpus. Now, the lists of terms are free from stop words and any numerical value which is not useful to reflect the quality of crawling. [6]

Step 5: Stemming is the process which is used for reducing inflected words to their stem, base or root form, which is generally a written form. For example, it will identify the string "stemming", "stemmer" and "stemmed" based on the word "stem". [6]

Step 6: In this step, the previous algorithm that is best first search algorithm is implemented. In which, adjacent matrix and heuristic score is calculated in order to build Best first search algorithm. It shows which document is linked with another document when the crawler is working. In this algorithm, graph is explored by expanding the most promising node according to some specific rules and this node is calculated from heuristic function. [5]

Step 7: In this the web pages are calculated which is having highest priority on the bases of keywords that come under that page related to domain.

Step 8: In this step, the proposed algorithm is shown in which is given in figure 3.2.

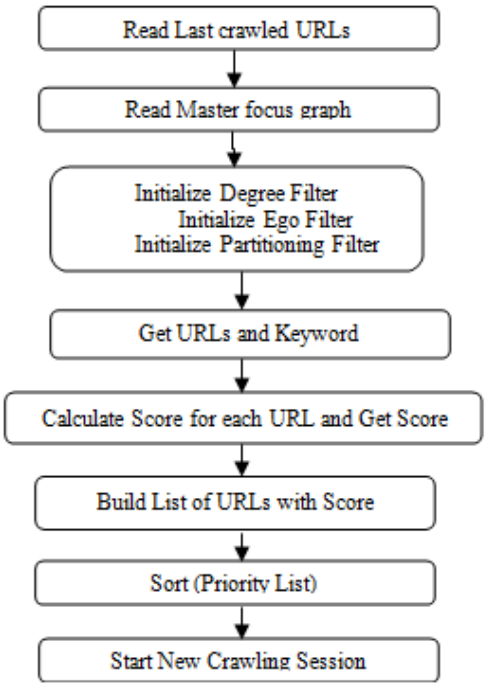


Figure 3.2: Proposed Architecture

The steps involved in the proposed work are as given below:

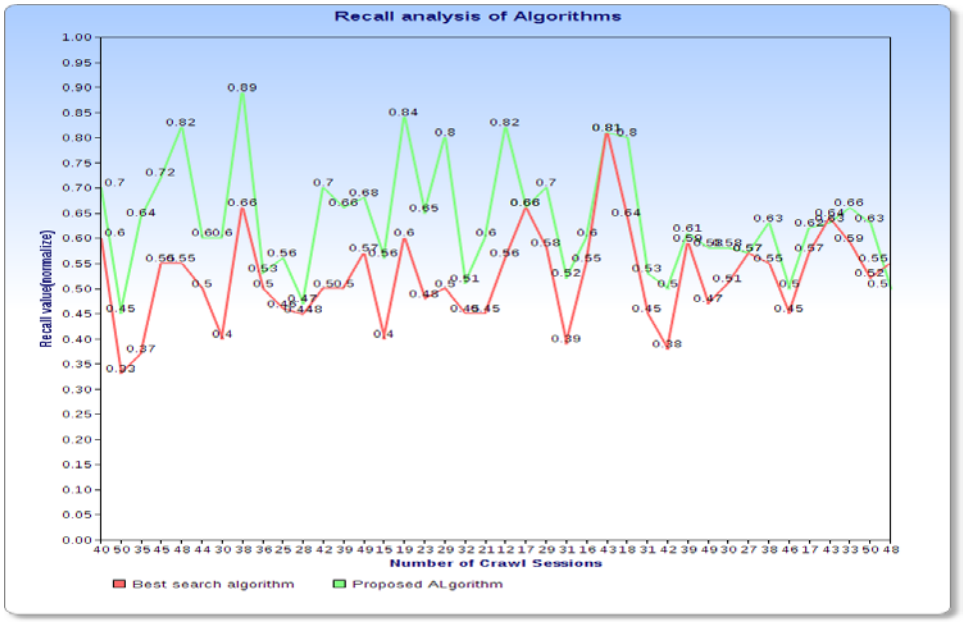
- In the first step of proposed algorithm, read the last crawl URLs. Building the master focus graph, their steps basically helps to build a focused domain graph from which all the crawlers' applications take seed as well as take reference from building a priority list for next crawl session.
- Next step is to build sub graphs and check the number of connections or edges between the nodes. There is more probability that would choose their path.
- If there is more number of connections/edges, this means that the crawler is most probably close to the sports or politics page having under network of these two domains only.
- The three filters degree, ego and partition are used to building sub graph, all these filters calculates score by traversing master graph depth search mechanism.
- Degree filter filters only those nodes that that have connection or at least a degree of 2. It will consider only that pages that is having degree 2 and discard other pages with degree except than 2.
- The partition filter finds the sub graph (connections/edges) based on the 'URL' text itself that the 'source node'.
- Next the ego filter creates a sub graph that helps to identify the underlining links to crawl, which has the keywords or set of keywords from the web page that has been crawled and matched with master focus graph. Ego filter works with regular expression.
- The combined score is calculated from all the filters.
- After find the score from three filters, ranking is done by sorting for highest score URL, they become the feed for the next crawl session.

4. Results

The result is based on two algorithms that are Best first search and proposed algorithm (Filters based Algorithm) in terms of recall and precision analysis of values.

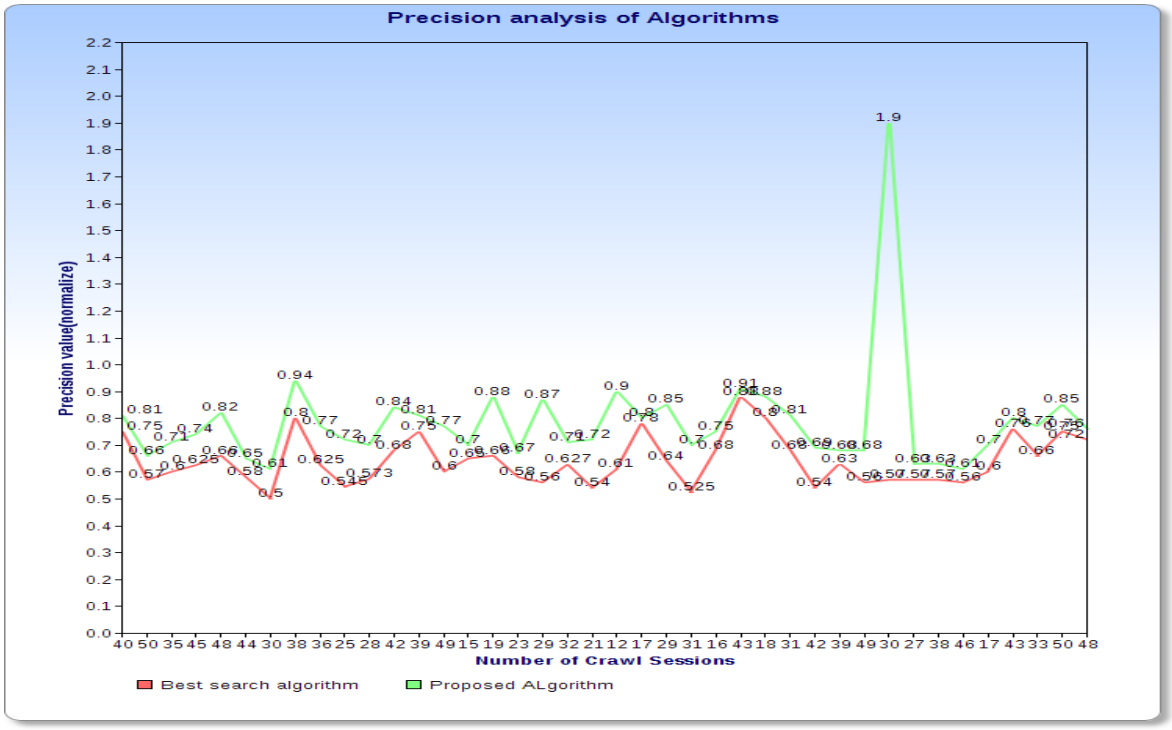
4.1 Recall analysis

The graph shows the values of recall, when crawler runs for each session. The proposed algorithm is able to search more relevant pages, as compared to previous algorithm.



4.2 Precision Analysis

The graph shows the values of Precision, when a crawler runs for each session. The proposed algorithm is able to search more relevant pages, as compared to previous algorithm.



5. Interpretation of Graph

The proposed focused crawler builds its corpus, which is specific to multi domains (sports, politics). Therefore, it is a model that works on the principle of selecting only those web documents for which as per proposed algorithm, it can gain information with respect to multi domains only. In this process it is intuitively reducing the uncertainty about the category of a document item being selected for crawling X provided by knowing the value of feature Y. Here item Y is the seed keywords or URLs or future hyperlinks or the titles. Since the ultimate goal of proposed algorithm or our focused crawler is to build a dataset that would provide a high information gain when used by a search engine or query engine, the selection of URLs and keywords is very important as it would lead to burning of less resources. As we are taking the advantage of highly optimized dictionary of terms related to sports (names of the sports, Name of Players, News related with sports) helps us in improving the recall and precision of our overall system.

It is apparent from the graph for recall analysis that the recall value varies from low to high which reflects the completeness or sensitivity of our proposed algorithm. The

recall value here means less number of craw jobs that are false negative in nature , or in simple words , crawling less number of web documents that were selected erroneously or those web URLs which were supposed to be rejected but got selected in URL crawl priority queue.

6. Conclusion

It can be concluded that after crawling, the number of relevant documents crawled and gather are more as compare to the previous algorithm and number of irrelevant documents are lea as compared to the previous algorithm. The three filters used in order to achieve this. The filters are used to build the score system that reflect the underlining/sub network source too. These filters use to make sub graph from the master graph, which retrieve more accurate results because by using master focus graph there is more probability that there is interconnections and closeness between word that is considered and the web page that is execute. The diversity within the topic of domain with respect to the crawled page is also good enough. The proposed algorithm saves resources in terms of time, bandwidth, etc. because only that pages are crawled which are related to particular domain due to maintenance of the master focus graph. The score calculated are more accurate as compare to the previous algorithm as the proposed algorithm providing a ranking system which is more accurate in helping the focus crawler to get domain based URLs only.

7. Future Scope

In this work, the master focus graph is maintained, from which full reference can be take to find the domain based keyword, based on which multiple scores are calculated from which a final score is extracted for running the URL for next crawl. This way the highly relevant document can be obtained. However, for future scope, a cross domain based reference master graph may be developed, based on the correlation of terms among related cross domains. Secondly, collaborative graph based clustering may be used to build graphs of URL for building next corpus of URL.

References

[1] S. Chakrabarti, M. Van den Berg, B. Dom, "*Focused crawling: a new approach to topic-specific web resource discovery*", in 8th international WWW Conference, May 1999.
[2] P.M.E. De Bra and R. D. J. Post, "*Information retrieval in the World Wide Web: making client-based searching feasible*", first conference in Computer Networks and ISDN Systems, Volume 27 Issue 2, Nov, 1994.
[3] M. Hersovici, M. Jacovi, Y. S. Maarek, D. Pelleg, M. Shtalhaim, and S. Ur, "*The shark-search algorithm. An application: tailored web site mapping*" Computer Networks and ISDN Systems, 30(17), 1998.
[4] Ali Pesaranghader, Ahmad Pesaranghader, Norwati Mustapha, Nurfadhlina Mohd Sharef, "*Improving Multi-term Topics Focused Crawling by Introducing Term Frequency-Information Content (TF-IC) Measure"*, 3rd International Conference on Research and Innovation in Information Systems, 2013, ISBN 978-1-4799-2486-8.
[5] Sunita Rawat, D.R.Patil, "*Efficient Focused based on Best First Search*", IEEE 3rd International Advance Computing Conference (IACC), 2013.
[6] Niran Angkawattanawit and Arnon Rungsawang, "*Learnable Crawling: An Efficient Approach to Topic-specific Web Resource Discovery*", Massive Information & Knowledge Engineering, 2002.
[7] Radhika Gupta and AP Gurpinder Kaur, "*Review of Domain Based Crawling System*", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.

Author Profile

Manpreet Kaur has done her B.tech degree from Guru Nanak dev Engineering College, Ludhiana in 2012. Currently, she is pursuing M.Tech degree from Chandigarh Engineering College, Landran (Mohali).

Yasmeen Kaur Dhaliwal has done her B.tech from Guru Gobind Singh College, Talwandi Sabbo in 2007 and M.Tech degree from Punjabi University in 2010. Currently, she is working as Assistant Professor in Chandigarh Engineering College, Landran (Mohali).