**Indonesian Journal of Electrical Engineering and Computer Science**
Vol. 13, No. 2, February 2019, pp. 492~498
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v13.i2.pp492-498


# Focused crawling from the basic approach to context aware notification architecture

**Venugopal Boppana, Sandhya P**
School of Computing Science and Engineering, Vellore Institute of Technology, Chennai Campus, India

**Article Info**

***Article history:***

Received Jul 7, 2018
Revised Oct 4, 2018
Accepted Nov 18, 2018

***Keywords:***

Complex event processing
Focused crawler
Topic specific crawler

**ABSTRACT**

The large and wide range of information has become a tough time for crawlers and search engines to extract related information. This paper discusses about focused crawlers also called as topic specific crawler and variations of focused crawlers leading to distributed architecture, i.e., context aware notification architecture. To get the relevant pages from a huge amount of information available in the internet we use the focused crawler. This can bring out the relevant pages for the given topic with less number of searches in a short time. Here the input to the focused crawler is a topic specified using exemplary documents, but not using the keywords. Focused crawlers avoid the searching of all the web documents instead it searches over the links that are relevant to the crawler boundary. The Focused crawling mechanism helps us to save CPU time to large extent to keep the crawl up-to-date.



***Corresponding Author:***

Venugopal Boppana,
School of Computing Science and Engineering,
Vellore Institute of Technology, Chennai Campus, Chennai, India.
Email: srees.boppana@gmail.com

## 1. INTRODUCTION

In recent days most of the latest information is available for us from the internet. But the greatest challenge is to get the relevant information for the given topic. This can also lead to extracting the irrelevant information from the web. This type of extraction, i.e., extracting both relevant and irrelevant data is done by the classical crawler. This lead to wastage of CPU time, memory and resources to large extent. The breadth first mechanism is followed by the classical crawler which searches all the links of a single parent. That possible links may consist of irrelevant data along with the relevant data.

To resolve the above challenges like time, space, resources and irrelevant data, topic specific crawler or focused crawlers are designed and introduced. These are much better than classical crawler in producing accurate data for the given topic. This topic specific crawler avoids the searching of the entire web, instead searches only specific area of the web. This crawler follows the mechanism of depth first search. The working of focused crawler is divided into two steps. In the first step irrelevant data is separated from the relevant data and the second step is selecting the seed page URL which helps in finding the next child nodes, i.e., next links for the relevant pages. The focused crawler helps in reducing the time to crawl, memory to store the crawled pages or to store the visited pages, decreases irrelevant data. This gives the great improvement over the classical crawler.

The classical focused crawlers and the learning focused crawlers are the two sub crawlers of the focused crawler. The classical focused crawlers are given with the predefined set of rules to pick the relevant pages for the given topic. Learning crawler updates the crawling link by learning from the training set. This training set is updated regularly.

## 2. THE CLASSIFICATION OF THE FOCUSED CRAWLER

Under the focused crawler we have two sub divisions (i) Learning focused crawler (ii) Classical focused crawler. Under the learning focused crawler, as shown in Figure 1, we have two sub division: (i) ANN based classifier (ii) Feedback method. Under classical focused crawler we have two sub divisions (i) Semantic crawler (ii) Social Semantic Crawler. Under the semantic crawler we have four sub divisions (i) Ontology and focused crawler model (ii) Context based approach for relevance (iii) Ontology based crawler (iv) FCA based approach. Under the social semantic crawler (i) Tag based approach crawling profile page (ii) Ontology based approach ontology web resources.

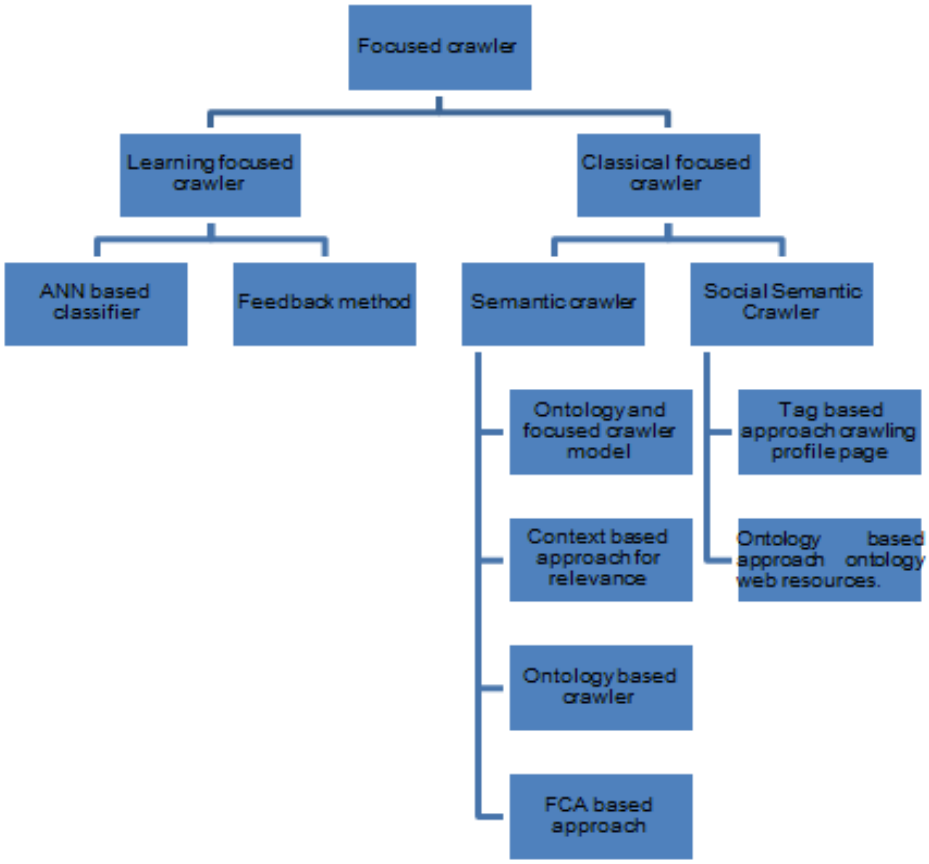


Figure 1. Classification of the focused crawler

## 3. LITERATURE SURVEY

Chakrabarti et al. [1] introduced the focused crawler to the world. The focused crawler first designed was based on the hypertext structure. The two important components in the working of the crawler are the classifier and the distiller, here we divide the relevant pages from the irrelevant pages by using the classifier and to find the seed URL we use the distiller, this seed URL leads us to other relevant pages, only based on the good seed URL we get good number of relevant links. This focused crawler has proven to give a better result than the classical crawler. The focused crawler is able to bring out more number of relevant pages when compared with the classical crawler. But the focused crawler will work as classical crawler if the seed URL is not selected accurately or if the training set is not sufficient.

In H. Zhang et al. [2] classification was done using the Artificial Neural Network (ANN), this paper used the ANN designed using the domain ontology. The methodology in this paper consists of three steps, the first step is data preparation, second step is training stage and the last stage is crawling stage. Here in the training stage ANN was used

In S. Chakrabarti et al. [3] other version of the learning focused was discussed in this paper. The methodology in this paper takes the two classifiers instead of a single classifier. The name of those two classifiers are critic classifier and apprentice classifier. The first classifier is used to collect the feedback and

the second classifier is used to train the basic classifier based on the feedback. Here the performance of getting the relevant pages is improved by adding other classifier to train the basic classier using the feedback obtained. The feature which are used to train the basic classifier by the second classifier are fully dependent features. This dependent feature greatly reduces the performance of the classifier.

The above three papers discussed about the learning focused crawler. Now we move to semantic focused crawlers which compare the meaning of pages visited with the search topic. If the meaning of the search topic matches with the page, then that page is considered as relevant page otherwise it is considered as irrelevant page. To compare the crawled page with the searched topic we use the ontology. The ontology maintains related words of the searched topic. The ontology widens the search area, which leads to more relevant pages and less irrelevant pages. Thus, using the ontology, we can easily compute the relevance of the visited page. Thus, the ontology plays an important role in relevance calculation. To extract the pages which are semantically matching with the search topic we can use the ontology with the focused crawler

In M. Ehrig et al. [4] gave the introduction and working of the focused crawler with the ontology. The paper has proposed a methodology containing couple of cycles. The first cycle is ontology cycle and second one is crawling cycle. The user query contains number of keywords, the relationship between these keywords is done in the first cycle. In the second cycle visited pages are collected based on the keywords given by user. In this paper they have adopted the breadth- first mechanism to crawl the relevant pages. But the disadvantage of this paper is, it can decide whether a particular page is relevant or irrelevant page only when complete page is downloaded. This leads to waste of resources.

Next, we shift our focus on to the social semantic focused crawlers. The social semantic focused crawlers use the social sites and social websites to get more relevant pages by learning the user profile and preferences. This working of focused crawling using the socio network has proven to give the better results. This crawling mechanism uses the knowledge of many expert people located in many places. As this crawling mechanism brings together knowledge of many experts to one place, thereby giving rise to many relevant pages for the searched topic.

To reduce the effort of user in searching the relevant page, tagging is introduced in the social network. This reduce the resource usage and brings out the more relevant pages than the irrelevant pages. A Social Semantic Focused Crawler combines both semantic knowledge and social network to get more relevant pages and links. The first approach combining the focused crawlers with social network and tagging was given by Z. Zhang et al. [5]. The focused crawling based on the profile page. In Nidhi Singh [6], showed the topic classification by using very minimum text which is available in URL. Instead of looking at the entire web page, just by text in URL we can classify the sentence based on the topic. This paper introduced online incremental learning algorithm to classify the URL.

A new traversal framework in focused crawling has been proposed by Siti Maimunah, Husni S Sastramihardja, Dwi H Widyantoro, Kuspriyanto [7] which increases the recall.As the conventional focused crawlers were only able to reach relevant web documents which connected directly which is not sufficient as there may exist web documents which are linked to identified relevant web documents.This can be achieved using this proposal. In Weng J, Lim E-P, Jiang J, He Q [8] proposes TwitterRank an extension to Page Rank algorithm.This algorithm measures the twitterers influence on topic-sensitivity.The proposed architecture performs topic distillation, constructing topic specific relationship network and finally providing ranks based on topic sensitivity. An Event Focused Crawling (EFC) architecture has been proposed by Farag, M.M. Gand E.A.Fox. [9].This crawler is used to retrieve highly relevant web pages which are similar to the selected seed URL's by the curator.This paper explains how focused crawler can be used to build an event model.

In Akyol, Mehmet Ali, et al. [10] discussed about a distributed architecture where distributed focused crawler and a distributed complex event processing are combined to identify the context of the users and notify them accordingly. The distributed focused crawler can be used to crawl the websites which are obtained from various data sources. Here distributed crawler is used to serve many users. The results of distributed crawler delivered to the users in based on their context.

## 4. FOCUSED CRAWLERS

The other names for the web crawler are bots, spider etc. These web crawlers form a structure of web pages and URL based on the user query. This software, based on the keywords in the user query searches for the URL and produces the relevant pages. The advanced and improved version of web crawler is the focused Crawlers. These focused crawlers based on the user search topic finds the seed URL and then from the seed URL, the crawler searches the relevant pages. The main aim of the focused crawler is to reduce the percentage of irrelevant pages with the total number of searched pages and increase the percentage of relevant pages with the total number of fetched pages. Under focused crawling we have two main divisions they are (i) Classic focused crawler (ii) Learning focused crawler, as shown in Table 1.

### 4.1. Classic focused crawler

The classical focused crawler is again divided into two crawlers, they are (i) Social Semantic Focused Crawler (ii) Semantic Focused Crawler. These are divided based on two criteria's first one is based on crawling area and second one based on the method followed to check the relevance of the fetched page.

The working of page relevance is almost same as computing a relevance of hypertext document. This crawler maintains the queue to collect all the fetched pages and URL. These pages are arranged based on the priority and ranking of the page. The name of this queue is priority queue. Here we use the page priority criterion to check the relevance of the page. The working of page priority criterion is similar to the distiller. The distiller can extract the relevant pages by detecting the good access point. Again, to get the good access point the crawler need to identify the good hypertext nodes.

Based on the application the crawler maintains the various priority queues. If the crawler finds the irrelevant page, then that link is not included in the queue. The crawler stops searching from that link and searches for the other link which leads to relevant pages. This the major difference between classical focused crawler and generic crawler. Many search engines use these genetic crawlers. After the crawler reaches the required number of relevant pages or if the time limit exceeds, crawler stops searching and return the result to the user.

### 4.2. Learning focused crawler

The second type of focused crawlers is Learning Crawler. These crawler work based on the training set. These take the feedback using the training set to update the crawling links which leads to more number of relevant pages. A group of sample pages related to the searched topic is taken as the training dataset. This training set helps in detecting the relevant and irrelevant pages. The various methods are followed by the learning crawlers. Some of them are Bayesian classifier, Hidden Markov Model. To compute the distance between crawled page and set of training pages, we can use context graphs.

### 4.3. Semantic and Social Semantic Focused Crawler

This section deals with the design of semantic and social semantic focused crawler. They crawl on different types of Web areas using different approaches. Focused Crawling based on Human Cognition (FCHC) crawling approach extracts the data related to relevant pages from the bookmarks given by the user. This maintains number of related or similar words for a single keyword. Now after the user has given the topic of search, then using the similar words the crawler can easily extract the web links of relevant data. These two crawling mechanisms work using two different patterns. The two patterns are Breadth-First Pattern (BFP) and Depth-First Pattern (DFP). The other variation of focused crawler is semantic focused crawler. This is also called as dynamic semantic relevance crawling (DSR). This arrange the pages visited in priority order.

### 4.4. Focused Crawler using Human Cognition (FCHC)

To produce the best results i.e. more number of relevant pages with minimum link search by the focused crawler, the choice of the seed URLs is very important as this seed url helps to find the other relevant links. Due to this efficient working of Focused crawling this can be applied with social media, here we can get large number of bookmarked pages based on the user interest. From the social media we can get the input from number of people with varied interest located at various place in the world.

From many years, the researchers are studying on "how to link the social media data with the web pages?" as this study can help to bring out more number of relevant pages with less time and resources. The main components of FCHC are:

***a) Selection of seed URL***

For every query given by the user the search engine produces n number of web URL from those URL, top priority URL are considered as the seed URLs which leads to many relevant URL. Selection of seed URL based on the topic given can bring large difference in the result of the crawler. A good seed URL helps the crawler to produce the best result. Crawler can select more than one seed URL, by this searching area can be wider than a narrow direction.

***b) Crawling area***

Following can be taken as crawling area, any site containing the data, site maintaining the bookmarking of the pages.

***c) Page relevance criterion***

While crawling the crawler matches the given topic with the pages visited to check whether it is relevant or irrelevant page.

***d) Page priority criterion***

A systematic search pattern motivated by human cognition is used as the priority criterion for the crawler. The two search patterns are Breadth first pattern (BFP) and Depth first pattern (DFP)

***e) Termination criterion:***

Based on the two conditions the crawler stops the crawling. The first condition is number of URLs to be crawled exceeds the limit and the second condition is till the priority queue is empty.

#### 4.4.1 FCHC Searching Patterns

Using both breadth first pattern (BFP) and depth first pattern (DFP) the pages can be searched to get the relevant pages. The flow of working in BFP is as follows, in this first all the users who tagged the seed page are placed in the queue, then the crawler starts visiting the pages from all the pages which are tagged by the user. This is done to get the resource of interest. Then after the crawler finds the relevant pages then those pages are stored in a queue. Most of the crawler work based on the BFP compared with the DFP.

There is a slight difference in parsing the pages in DFC, in this instead of visiting the pages from seed URL of all users at same time first one user is picked and from here parsing is done till the crawler reaches the resource of interest and then the crawler starting again the parsing from the other user till it reaches the resource of interest. This procedure continues till all the users are completed.

### 4.5. DSR based Semantic Focused Crawler

This DSR fetches the topic relevant pages from the particular area using the multithreading concept. To get more relevant pages on a given topic we can use domain ontology. Mostly for the educational purpose we can use the domain ontology to expand the topic.

#### 4.5.1 DSR based Semantic Focused Crawler Framework

To design the efficient DSR we need the following components they are of domain ontology, local database, priority queue, and the proposed multithreaded Semantic Focused Crawler. SFC (Semantic Focussed Crawler) picks the web page that can direct the crawler to many other relevant pages. Generally, SFC selects top rated URL. We get this top-rated URL from the priority queue. Here to parse the web, number of parallel threads are created, by this we can get number of hyperlinks at same time. These hyperlinks are added to the queue. These hyperlinks are used to parse the web to get more of number relevant web pages. The sequence of parsing of the URL also plays an important this order of parsing can be known from the priority queue. The crawler should avoid the visiting same old page number of times. To avoid this situation another queue is maintained which stores the pages visited. The hyperlinks of the relevant pages are stored in separate database to be used for later purpose.

### 4.6. Comprehensive Traversal focused Crawler

The conventional focused crawlers follow the top down approach in order to get the topic specific web documents which is useful when there is only one link which is topically specific. But if the root node of web document consists of another relevant document linked to this node the crawler cannot go back and because of this we will get low recall.To improve recall this framework has been proposed improves the recall of the crawling in an impressive manner. To improve this a lexicon list is prepared where document relevance can be assessed from the local ontology.

### 4.7. Event Focused Crawler

This is an architecture where event modelling can be done using Event Focused Crawler. Based on the context and type, the events can be recognised and represented. The context here is nothing but *when, where*. The type means *what*. This can be used to prepare list of seed URL's based on the events. Using the event model analysis can be done on event collections.

### 4.8. The Context Focused Crawler

This helps user to query the search engine for page that has a link with a particular document. This mechanism is possible is Context Focused Crawler (CFC). This query helps to construct a context graph of pages which are at minimum distance from the URL of the page given by the user. This minimum distance is decided based on the application. Here the minimum distance is the number of links used to reach the relevant page from the page URL given by the user. This constructed structure can be used in the training of the classifier. Then the classifier divides the pages according to the topic. This division is based on the distance traversed by the crawler to reach the target document.

They are two stages in context focused crawler:

1) An initialization phase: In this context graph and associated classifiers are constructed for every seed documents
2) A crawling phase: In this search engine by using classifier traverse to reach the relevant document. Based upon these links updation in context graph are done.

Table 1. Summary of Crawlers

| Crawler Name | Description |
|---|---|
| Classic focused crawler | This crawler maintains the queue to collect all the fetched pages and URL. |
| Learning focused crawler | Take the feedback using the training set to update the crawling links which leads to more number of relevant pages |
| Focused Crawling based on Human Cognition (FCHC) crawling | This approach extracts the data related to relevant pages from the bookmarks given by the user. |
| Focused Crawler using Human Cognition (FCHC) | Extracts large number of bookmarked pages based on the user interest |
| FCHC Searching Patterns | Uses both breadth first pattern (BFP) and depth first pattern (DFP) to search the relevant pages |
| DSR based Semantic Focused Crawler | This fetches the topic relevant pages from the particular area using the multithreading concept. This uses the domain ontology |
| Comprehensive Traversal focused Crawler | This uses top down approach in order to get the topic specific web documents |
| Event Focused Crawler | To extract the relevant pages, event modelling analysis is used. |
| The Context Focused Crawler | Helps the search engine to query for page that has a link with a particular document. This query helps to construct a context graph of pages |

## 5. A FOCUSED CRAWLER IN CONTEXT AWARE NOTIFICATION ARCHITECTURE

In pull-based system user may miss some of the important information or cannot get the updated information. This can be resolved by using the push-based notification technique. This can be achieved by introducing the focused crawler in context aware notification. Using this technique, the user can receive the latest information based on the contexts specified by the user. The biggest advantage of the push-based notification is it helps the users to get the latest information by avoiding continuous querying by the user. Here the user first need to specify his interested topic and context, based on the interest given by the user focused crawler send the notification of latest information about that particular topic. This also send the notification to the user based on the context i.e. location, time etc.

The context can be divided into two categories first one is external and second is internal. We can get the information about place, temperature, light, sound, and air pressure by using the sensors. This type of information comes under external context. The internal context are the user preferences. To achieve the better results, the distributed architecture need to be designed to traverse the required URL and to send the notification to the user based on both internal and external context. The user can receive the notification via SMS, chat-bot messages, email. The framework need to be designed such that it should allow the user to specify the context to receive the information. The user can specify the time, location, notification method to receive the information. The Figure 2 shows the architecture of conceptual framework.

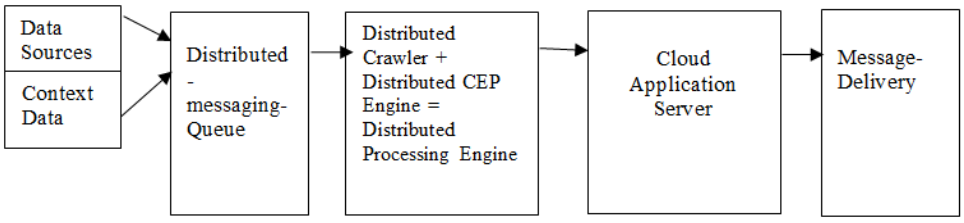


Figure 2. Architecture of Conceptual framework

The various data sources are facebook, twitter, websites and context data considered is time, location, keywords, preferences of the user. Now context data and data sources are maintained on distributed messaging queue. Now based on the information and context specified by the user, distributed processing engine using the data stored in cloud application server, send the notifications to the users using SMS, email or chat bot. This distributed processing engine consists of distributed crawler and distributed cep engine.

## 6. CONCLUSION

The focused crawler solved many problems of the generic crawler and helped to get the more relevant pages with minimum number of traversals. We have given a overview of many versions of the focused crawler by specifying its advantages and disadvantages. The focused crawler had brought good and great change in searching for the given user query. To search the complete web to get the relevant pages more than one focused crawler can be used. This gives less number of irrelevant pages and more number relevant pages based on the user query. According to the above discussed focused crawler used tags and ontology to get relevant pages and also to expand the area of searching. By using more efficient tagging method the performance of focused crawler can be improved. We can also include the context aware notification in focused crawler to bring out more relevant pages. We can improve the focused crawler performance by machine learning algorithms. This helps to compare the web pages with content posted by the user. To process this content we may use the text mining algorithm like feature selection.

## REFERENCES

[1] S. Chakrabarti, M. Berg, and B. Dom. Focused Crawling: A New Approach to Topic-specific Web Resource Discovery.*Journal of Computer Network*.1999; 31(11-16) :1623-1640.
[2] Z. H. Tao, K. B. Yeong, K. H. Gee. An ontology-based approach to learnable focused crawling. *Journal of Information Science*.2008; 178(23):4512-4522.
[3] S. Chakrabarti, K. Punera, and M. Subramanyam. *Accelerated Focused Crawling through Online Relevance Feedback.* In Proceedings of 11th International conference on World Wide Web.2002; 148-159.
[4] M. Ehrig, and A. Maedche. *Ontology-focused crawling of web documents.* In proceedings of ACM symposium on applied computing, pp. 1174-1178, 2003
[5] Z. Zhang, O. Nasraoui and R. Zwol. *Exploiting Tags and Social Profiles to Improve Focused Crawling.* International Joint Conferences on Web Intelligence and Intelligent Agent Technology, pp. 136-139, 2009
[6] Singh, Nidhi, et al. *Large scale url-based classification using online incremental learning*. *11th* International Conference on Machine Learning and Applications (ICMLA), 2012. Vol. 2. IEEE, 2012.
[7] Siti Maimunah, Husni S Sastramihardja, Dwi H Widyantoro, Kuspriyanto. CT-FC: more Comprehensive Traversal Focused Crawler. *TELKOMNIKA Telecommunication, Computing, Electronics and Control* Vol. 10, No. 1, March 2012: 189 – 198. ISSN: 1693-6930.
[8] Weng J,Lim E-P,Jiang J,He Q.*TwitterRank: finding topic-sensitive influential twitterers*. Proceedings of the third ACM international conference on Web search and data mining. New York, USA. 2010; 261-270.
[9] Farag, M.M.G. and E.A.Fox. Building and archiving event web collections: A focused crawler approach. *in Bulletin of IEEE Technical Committee on DigitalLibraries.* 2015; p.1-2.
[10] Akyol, Mehmet Ali, et al.*A Context Aware Notification Architecture Based on Distributed Focused Crawling in the Big Data Era*. European, Mediterranean, and Middle Eastern Conference on Information Systems.Springer, Cham, 2017.