International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 09 | Dec-2015 www.irjet.net p-ISSN: 2395-0072

An Overview of Approaches Used In Focused Crawlers

Parigha Suryawanshi [1], D.V.Patil [2]

Student, Department of Computer Engineering, GES RH Sapat COE, Maharashtra, India
Associate Professor, Department of Computer Engineering, GES RH Sapat COE, Maharashtra, India

---

Abstract - ***Web is a repository n where there is variety of information available provided by millions of web content providers. Numerous WebPages are added to web every day and the content keeps changing. Search engines are used to mine this information and the most important part of search engine is a web crawler also known as web spider. A web crawler basically is software that crawls or browses the WebPages in the World Wide Web. There are many types of crawlers having different methods of crawling like parallel crawler, distributed crawler, focused crawler, parallel crawler, and incremental crawler. In recent years, focused crawling has attracted considerable interest in research due to the increasing need of digital libraries and domain-specific search engines. This paper reviews different researches done in focused crawler which are also known as topic specific web crawler.***

Key Words: *Crawler, Search Engine, Focused Crawler*

1. INTRODUCTION

A search engine has become an important source for mining the data in the World Wide Web (WWW). Since the web crawler is the main part of the search engine it needs to browse WebPages that are topic specific. A web crawler is basically a software or program which browses the internet and collects data in a repository. In process of crawling the web crawler gathers WebPages from the web and stores them in a proper way so that the search engine can retrieve them quickly and efficiently.

A web crawler starts with a URL also called as seeds which are stored in the crawler frontier. Then it identifies the hyperlinks while parsing the web pages and adds them to the list of URLs that already exists and the collected data by crawler is sent to storage. This process of crawling depends on the policies defined for the crawler. The general architecture of crawler is shown in figure 1. The frontier consists of the list of unvisited URLs. The crawler fetches a URL from the frontier which has the list of unvisited URLs. The page which corresponds to that URL is fetched from the Web and the unvisited URLs from that page are added back to the frontier. The process of retrieving and extracting the URL goes on till the frontier is empty or some other situation causes it to stop [1-3].The main job of the page fetcher is to fetch the pages from World Wide Web corresponding to the URLs which has been retrieved from the crawler frontier. For that purpose, the page fetcher requires a HTTP client for sending the HTTP request and to read the response. Web Repository stores the web pages in the database which it receives from a crawler. All other multimedia and document types are avoided by the crawler .It stores the browsed pages as different files and the storage manager stores the updated version of each page fetched by the crawler.

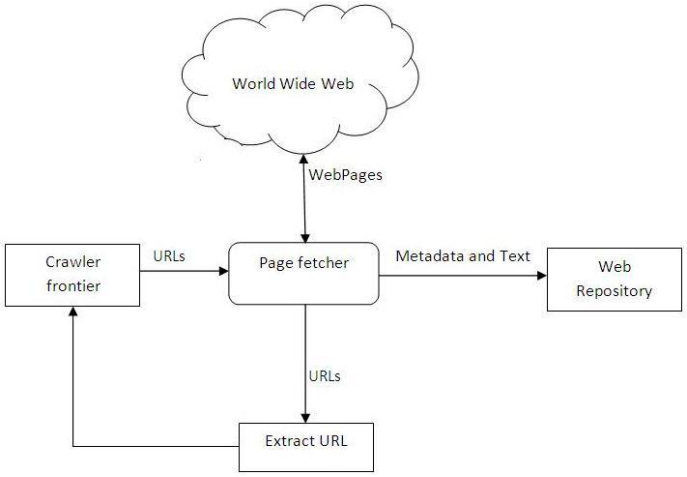

Fig -1: General Architecture of Web Crawler

Generic crawlers do not specialize in specific areas. A traditional crawler periodically crawls the URLs that are previously crawled and replaces the old documents with the newly downloaded documents to refresh its collection. On the contrary, an incremental crawler refreshes the already existing collection of pages gradually by visiting them frequently. This is based upon an estimation of the rate at how often pages change. It also replaces old and less important pages by new and more relevant pages. It resolves the problem of freshness of the data. The advantage of incremental crawler is that only valuable data is provided to the user [4-8].

Focused Crawler is a web crawler for fetching web pages that are related to a specific area of interest. It collects the documents that are focused and relevant to a given topic. It is called as a Topic Crawler because of the way it works. The focused crawler determines the relevance of the document before crawling the page [9]. It estimates if the given page is relevant to a particular topic and how to proceeds. The main advantage of this kind of crawler is that it requires less hardware resources [10].

2. FOCUSED CRAWLER SYSTEMS

Soumen Chakrabarti [10] introduced the system of Focused Crawler in 1999. In this system two hypertext mining programs a classifier and a distiller were used that guided the crawler. Classifier evaluates the relevance of a hypertext document with respect to the topic and the distiller identifies hypertext nodes that are access points to numerous relevant pages within a less number of links. Focused crawler fetch topic specific pages steadily while the traditional crawling quickly loses its way even if it starts from the same root set. It is capable of exploiting valuable resources that are away from the root set. Focused crawling was found very effective for fetching high quality web documents on the specified topics and also it required less hardware resources.

2.1 Context model based focused web crawler

The main challenge in focused crawling is fetching topic relevant data in less time but also crawling through the path that eventually yield large collection of relevant pages. To solve this problem M. Diligenti et al [11] presented a focused crawling algorithm which uses a context graph within which topic relevant pages available on the web occur.
The TF-IDF score n(w) of a phrase w is computed using the following function.

$$n(w) = f\,d(w)/f\,dmax * log\ N/f(w)$$

where *fd(w)* is the number of occurrences of w in a document d, *fdmax* is the maximum number of occurrences of a phrase in a document d, N is the count of documents in the reference corpus and *f (w)* is the count of documents in the corpus where the phrase w occurs at least once.

It captures link hierarchies in which relevant pages occur, and it models the content in the documents which occur frequently within the relevant pages. This algorithm also increased the partial reverse crawling capability of search engines. This approach that models the links and content in the documents which are relevant improved the performance and so the topic specific content can be found. The limitation of this approach was its need for reverse links from a known search engine for the seed set documents links.

2.2 Document Object Model based approach

Document has always been a fundamental unit in traditional Information Retrieval .The documents are generally static HTML files. Soumen Chakrabarti et al [12] presented a Document Object Model (DOM) based fine grained view of hypertext and an enhanced topic distillation algorithm that analyze hyperlinks within pages, trees of markup tags that constitute HTML pages and the text. It therefore identifies the text and hyperlinks that are highly related to the specified query and those sub-trees get priority during the crawling process. This integrated and propagated micro-hub scores at a level which was determined by relevance of text to the query and the analyzed graph structure. This approach was found to reduce topic drift and capable of identifying hub regions relevant to the search query.

A new technique of using two classifiers was introduced by Soumen Chakrabarti et al [13]. The role of assigning priorities is now assigned to a new learner called the apprentice was given the job of assigning priorities to URLs in the crawl frontier which were not visited. A baseline classifier was used to generate training set for the apprentice which uses the features derived from the Document Object Model (DOM). These improvements made focused crawler to assign better priorities to URLs in the crawl frontier which lead to fetching of relevant pages relevant to topic at higher rate. Also there was no need to train the system manually to seek the path having relevant pages.

2.3 Learnable focused crawlers

Some site's contents typically are kept in databases and less exploited. Since the volume of such hidden data was growing, there was requirement of some techniques that would help users to exploit this hidden data [14]. A new crawling strategy was introduced by Luciano Barbosa [15] to automatically discover hidden-Web databases. This Crawler performs a broad search by focusing the search on specified topic efficiently by learning to identify promising links. The crawler uses two classifiers the page and the link classifiers to guide its search. And the third classifier, the form classifier filters out useless forms. The page classifier is trained to classify pages according to topics specified. When the crawler fetches a page links are extracted from it. The link classifier examines links which are extracted from topic specific pages and adds them to the crawler frontier according to their priority. The link classifier is trained to identify links that can to lead to pages which consist of searchable form interfaces. But this approach requires considerable manual tuning, also

appropriate features needs to be selected and the link classifier has to be created.

A new framework was proposed by Luciano Barbosa et al [16] where crawlers could learn patterns of relevant links automatically and accordingly adapt their focus while crawling the web. This method reduced work of manual setup and tuning which was a major drawback of FFC crawlers. Adaptive learning strategy effectively balances the exploration of acquired knowledge with the exploitation of links with unknown patterns, making it robust and able to correct biases introduced in the learning process. Since this crawler learn from scratch ,it is able to get harvest rates that is equal or even more than the manually configured crawlers, hence the ACHE framework reduces the effort to configure a crawler manually.

The main characteristic of focused crawler is to collect topic relevant pages. The previous crawling experience can help crawler to build knowledge base and learn from it, so that it can improve its performance. Niran et al [17] presented an algorithm that built knowledge bases using seed URLs, keywords and URL prediction. These knowledge bases helps crawler to learn and produce the result in more efficient way. The knowledge bases are incrementally built from the log of previous crawling. Seed URLs allow the crawler to collect as many relevant web pages as possible. Keywords support the crawler to recognize appropriate documents. URL prediction enables the crawler to predict the relevancy of the content of unvisited URLs .Qingyang Xu et al [18] introduced a new general framework for focused crawler that is based on "relational subset discovery". To depict the relevance within the pages that are not visited in the crawl frontier predicates are used ,after that first order classification rules are introduced using subgroup discovery technique. Then the learned relational rules guide crawler with sufficient support and confidence.

A new approach for predicting the links that lead to relevant pages based on a Hidden Markov Model (HMM) was presented by Hongyu Liu et al [19]. The system includes three stages: user data collection, user modeling using sequential pattern learning, and focused crawling. Initially web pages are collected while browsing through web. Then these WebPages are clustered, and then hierarchical linkage pattern within pages from different clusters is then used to learn sequence of pages that lead to target pages. The Hidden Markov Model (HMM) is used for learning process. During the crawling process the priority of links are decided on basis of a learned estimate of how likely the page will lead to a target page. The performance was compared with Context-Graph crawling during experiments it was found that this approach performed better than Context-Graph crawling.

Deep web contents are the information content that cannot be indexed by search engines as they stay behind searchable web interfaces. Feng Zhao et al [20] presented framework with two stages called Smart Crawler, for achieving efficient harvesting deep web interfaces. In initial stage, crawler does the site-based search for root pages using search engines and avoids visiting more pages. And for more accurate results for a focused crawl this system does ranking of websites and gives priority to more relevant ones for the specified topic. In the later stage it uses adaptive link ranking to exploit more relevant links in less time. A link tree data structure is also designed to get wider coverage for a website. Experimental results showed that this system efficiently retrieve deep-web interfaces achieving better harvest rates than other crawlers.

2.4 Ontology or semantics based focused web crawler

Keyword driven crawling, that can also decide relevancy of web pages to the topic can certainly improve the performance of a focused crawler .Keyword Focused Web Crawler presented by Gunjan H. Agre et al [21] which uses keywords to decide the relevancy while crawling through web and it also uses Ontology concepts to improve crawler's performance. It extracts URLs for web pages that consists of search keyword in their content and considers those pages only and not other irrelevant web pages. It gives more optimality as compared to traditional web crawler and enhances the search. This technique reduces the number of extracted web pages thus it takes less turn-around time for crawling process.

Crawlers generally use the senses offered by lexical database to recognize the web pages relevant to the search query. Combining text and link analysis approach was presented for focused crawling by G. Almpanidis et al [22].In this system a latent semantic indexing classifier was developed which combine link analysis with text so as to fetch and index topic-specific web documents. Advantage in this method is that it does not depend on historical information like previous crawl and index of the web, or any existing search services.

A technique was presented by Prashant Dahiwale et al [23] which use the semantics of URL and anchor text to check the relevancy. An algorithm called Anchor Word Processing was used in this technique. This approach in focused web crawler predicts similarity before downloading the web page. Hence it saves the bandwidth of downloading each unvisited page. Also the document quality of related links was found to be good.

.

2.5 Genetic algorithm based approach

Conventionally, focused crawlers adopt the Vector Space Model and confined search algorithms which find topic

related pages but with low precision value. Mostly the recall value is also low, since the crawler exploits a restricted sub-graph of the web that is nearby the seed URLs, and avoids other relevant pages on the exterior of the sub-graph. A Novel Hybrid Focused Crawling Algorithm was introduced by Yuxin Chen [24] that is based on Genetic Programming (GP) and meta-search. This approach combines an inductive machine learning algorithm through GP and meta-search technology to improve focused crawling performance. It was found that the GP framework can be used to discover better similarity functions, which can be used by classifier to achieve better results. The meta-search technique combines the results of multiple search engines to improve the efficiency of web search.

The size of web grows day by day which increases the necessity of domain specific search engines. Chain Singh et al [25] also used genetic algorithm to expand initial keywords for focused crawling. They found that this technique could provide high quality domain specific results than the traditional focused crawling techniques. Also a global search algorithm was incorporated into this technique. The main objective of this method is to expand the keyword set for the focused crawling. These method made searching easy and more relevant documents were found. According to their experiments, focused crawler assemble the collections with high efficiency than the general best-first crawler .The results of experiment conducted by them stated that average document's relevance was increased by up to 50%.

3. CONCLUSIONS

Focused crawling is process which goes on harvesting links which are relevant to the specified topic and discards other irrelevant links. A number approaches have been presented in last few years to improve focused crawling strategies. This paper describes some of those approaches for focused crawling. Apparently all crawling approaches certainly have their advantages and some drawbacks. So they can be used for the purpose of developing new approaches by considering their strengths and drawbacks.

REFERENCES

[1] "Modern Information Retrieval", book by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.
[2] "Web Data Mining" ,book by Bing Liu.
[3] Sanjeev Dhawan, Vinod ,"An Approach for Fetching User Relevant Pages Using Backlinks: A Review", IJARCSSE 2013
[4] "Mining the Web: Discovering Knowledge from Hypertext Data" ,book by Soumen Chakrabarti.
[5] "An introduction to Information Retrieval" ,book by Christopher Manning.
[6] "Information Retrieval", book by David A. Grossman and Ophir Frieder.
[7] Trupti V. Udapure, Ravindra D. Kale, Rajesh C. Dharmik, "Study of Web Crawler and its Different Types", IOSR Journal of Computer Engineering 2014.
[8] Hongyu Liu, Evangelos Milio1, Jeannette Janssen, "Probabilistic Models for Focused Web Crawling", ACM 2004.
[9]Donna Bergmark, Carl Lagoze, and Alex Sbityakov, "Focused Crawls, Tunneling, and Digital Libraries", Cornell Digital Library Research Group 2002.
[10] Soumen Chakrabarti, Martin Van den Berg, and Byron Dom, "Focused crawling: a new approach to topic-specific web resource Discovery", Computer Networks, 1999.
[11]M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles and M. Gori, "Focused Crawling Using Context Graphs", in Proceedings of VLDB 2000, pages 527-534.
[12] S. Chakrabarti, M. Joshi, and V. Tawde , "Enhanced Topic Distillation using Text, Markup Tags and Hyperlinks", in Proceedings of the ACM SIGIR, 2001.
[13]S. Chakrabarti, K. Punera, and M. Subramanyam, "Accelerated focused crawling through online relevance Feedback" In Proceedings of WWW, pages 148–159,2002.
[14]Mukesh Kumar, Renu Vig,"Learnable Focused Meta Crawling Through Web",Elsevier (ICCCS) 2012.
[15] L. Barbosa and J. Freire., "Searching for Hidden-Web Databases", In Proceedings of WebDB, pages 1–6, 2005.
[16] Luciano Barbosa and Juliana Freire, "An adaptive crawler for locating hidden-web entry points",In Proceedings of the 16th international conference on World Wide Web, pages 441–450.ACM, 2007.
[17]Niran Angkawattanawit and Arnon Rungsawang,"Learnable Crawling: An Efficient Approach to Topic-specific Web Resource Discovery",2002.
[18]Qingyang Xu , Wanli Zuo, "First-order Focused Crawling", International World Wide Web Conferences 2007.
[19] Hongyu Liu, Jeannette Janssen, Evangelos Milios "Using HMM to learn user browsing patterns for focused Web crawling" ,Elsevire Data & Knowledge Engineering 2006.
[20] Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang, Hai Jin, "SmartCrawler: A Two-stage Crawler for Efficiently Harvesting Deep-Web Interfaces", IEEE Transactions on Services Computing 2015.
[21]Gunjan H. Agre, Nikita V. Mahajan, "Keyword Focused Web Crawler", IEEE sponsored ICECS 2015.
[22]G. Almpanidis, C. Kotropoulos, I. Pitas, "Combining text and link analysis for focused crawling—An application for vertical search engines" ,Elsevire Information Systems 2007.
[23]Prashant Dahiwale, M.M.Raghuwanshi, Latesh Malik, "Design of improved focused web crawler by analyzing semantic nature of URL and anchor text", Industrial and Information Systems , IEEE conference 2014.
[24]Yuxin Chen, Edward A. Fox et. Al, "A Novel Hybrid Focused Crawling Algorithm to Build Domain-Specific Collections",Virginia Polytechnic Institute & State University Blacksburg, VA, USA pp- 85, 2007.

[25] Chain Singh, Ashish Kr. Luhach, Amitesh Kumar, "Improving Focused Crawling with Genetic Algorithms", International Journal of Computer Applications , March 2013.