A Survey on Web Crawling Algorithms Strategies

Chain Singh1, Kuldeep Singh2, hansraj
DCE, Gurgaon1,3 Research scholar, IIT-BHU, Varanasi2
Email: cschhoker@gmail.com1, ksbhan@gmail.com2, *hansrajy@gmail.com*

Abstract: Availability of huge amount of useful information and is expanding rapidly. Search engine play main roll. Web crawlers are one of the most key components in search engines and their optimization would have a great effect on refining the searching efficiency. Crawlers enable the process by following the hyperlinks in Web pages to automatically download a fractional snapshot of the Web. Crawling algorithms are thus key component in selecting the pages that satisfies the users' needs. This paper review researches on web crawling algorithms strategies used on searching.

Index Terms: web crawler, web algorithms

1. INTRODUCTION

A Web crawler is a key component inside a search engine. It can traverse the Web space by following Web page's hyperlinks and storing the downloaded Web documents in local repositories that will later be indexed and used to respond to the user's queries efficiently. A crawler is a program that automatically collects Web pages to create a local index and /or local collection of web pages. However, with the huge size and explosive growth of the Web, it becomes more and more difficult for search engines to provide effective services to end-users. Moreover, such a large collection often returns thousands of result documents in response to a single query. It is impossible for major search engines to update their collections to meet such rapid growth. As a result, end-users often find the information provided by major search engines not comprehensive or out-of date.[8] Web crawlers are programs which traverse through the web searching for the relevant information [1] using algorithms that narrow down the search by finding out the most closer and relevant information. This process is iterative, as long the results are in closed proximity of user's interest. The algorithm determine relevancy based on the factors such as frequency and location of

Keywords. Search engines use algorithms which sorts rank the result in the order of authority that is closer to the user's query. Many algorithms are is in use - . Breadth-First search, Best-First search, Graphic Context algorithm, Fish search, Shark search, Genetic Algorithm The search engine techniques may become useless or junky if the information it draws are not attracting users, especially if the malicious user who are trying to attract more traffic in to their site by embedding the most used keywords invisibly in to their site. The challenges are relevancy, robustness and the ability to download large number of pages.

2. FUNDAMENTALS OF SEARCH ENGINE

Search engines are based on a centralized architecture that relies on a set of key components, namely Crawler, Indexer and Searcher.

DEFINITION:-

2.1 ***Crawler:-*** A crawler is a module aggregating data from the World Wide Web in order to make them searchable. Several heuristics and algorithms exists for crawling, and based upon following links.

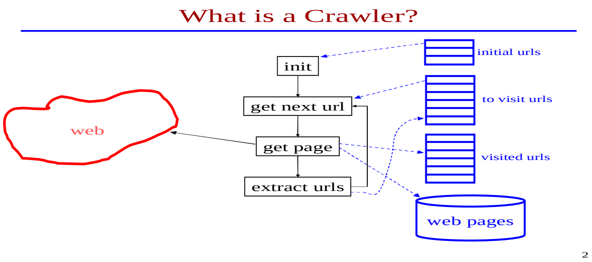
2.2 ***Indexer****:-* A module that takes a collection of documents or data and builds a searchable index from them.

2.3 ***Searcher****:-* The searcher is working on the output files from the indexer. The searcher accepts user queries, runs them over the index, and returns computed search results to issuer.[1]

Where the Web a static collection of pages we would have little long term use for crawling. Once pages had been fetched to a repository (like a search engine's database), there would be no further need for crawling. However, the Web is a dynamic entity with subspaces evolving at differing and often rapid rates. Hence there is a continual need for crawlers to help applications stay current as new pages are added and old ones are deleted, moved or modified.

3. WHY DO WE NEED A WEB CRAWLER?

Following are some reasons to use a web crawler:-
To maintain mirror sites for popular Web sites.
To test web pages and links for valid syntax and structure.
To monitor sites to see when their structure or contents change.
To search for copyright infringements.
To generate a special purpose index. For example, one that has some meaning of the content stored in multimedia files on the Web [20].



Defining the behavior of a Web crawler is the outcome of a combination of below mentioned **strategies**.
Selecting the better algorithm to decide which page is download. Strategizing how to re-visit pages to check for updates.Strategizing how to avoid overloading websites [20].

3.1 How to Re-Visit Web Pages

The optimum method to re-visit the web and maintain average freshness high of web page is to ignore the pages that change too often. The approaches could be:

1. Re-visiting all pages in the collection with the same frequency, regardless of their rates of change is called periodic crawler
2 Re-visiting more often the pages that change more frequently is called incremental crawler [20].

3.2 Selecting the Right Algorithm

While selecting the search algorithm for the web crawler an implementer should keep in mind that algorithm must make sure that web pages are chosen depending upon their importance. The rank of a web page lies in its popularity in terms of links or visits, or even it's URL.

4 CRAWLING ALGORITHMS:

We now discuss a number of crawling algorithms that are suggested in the literature. The difference is in the heuristics they use to score the unvisited URLs with some algorithms adapting and tuning their parameters before or during the crawl [19]. Crawlers are designed for different purposes and can be divided into two major categories.

1. Breadth-First search.
2. Best-First search.
3. Graphic Context algorithm.
4. Fish search.
5. Shark search.
6. Tunneling.
7. Info Spider.
8. Genetic Algorithm.

4.1 Breadth-First Search Crawler

The first generation of crawlers on which most of the web search engines are based rely heavily on traditional graph algorithms, such as depth-first or breadth-first traversal, to index the web. A set of URLs are used as a seed set, and the algorithm repeatedly follows hyperlinks down to other documents. Since the ultimate goal of the crawl is to cover the whole web. A standard crawler follows link, typically applying a breadth first strategy. If the crawler starts from a document which is I steps from a target document, all the documents that are up to I - 1 steps from the starting document must be downloaded before the crawler hits the target.

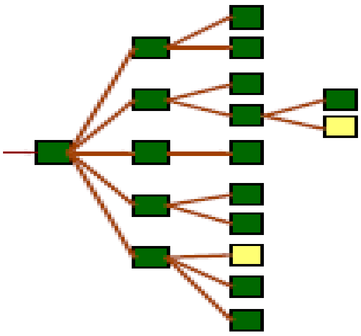


Fig. 2

4.2 Best-First

A rule about what it is 'best's defined (in most cases a score or rank). When a link has to be selected from the frontier to be fetched that rule is applied. In most cases a classifier algorithm is applied such as Naive Bayes, Cosine Similarity. A best first search is performed by popping the next page to analyze from the head of the queue. This strategy ensures that the crawler preferentially pursues promising crawl paths.

4.3 Graphic Context Algorithm.

The graphic context is an algorithm that is explained in the figure is shown the idea of this algorithm. Some relevant pages are found manually and they will be the layer 0 (from this point L0). With the help of a search engine the back-links to those relevant pages are found and they become the L1, again with a search engine are found the back-links of the L1 and they become the L2. This is repeated recursively as many times as we wish (L0 L1 L2...). After the whole process a Context Graph is obtained. It gives an idea of which pages point/lead to the deeper layers in the circle, the closer is to the center the better the page is. This is because if a web-page that is being processed is on L1 means that its out-links (or at least few) will lead to L0 (the related page), thus those out-links should be fetched as soon as possible because they seem really expected.[12]

4.4 Fish Search

System is one of the earliest approaches to ordering the crawl frontier (for example, through a priority queue of URLs). The system is query driven.
Starting from a set of seeds page, it considers only those pages that have content matching a given query (expressed as a keyword query or a regular expression) and their neighborhoods (pages pointed to by these matched pages)[19].

Metaphor: School of fish. If food (relevant info) is found then reproduce and continue looking. If not (no relevant info) or polluted water (poor bandwidth), then die. Search only to fixed depth from seed.

4.5 Shark Search

System is an improvement over fish search. It uses a weighting method of term frequency (TF) and inverse document frequency (IDF) along with the cosine measure to determine page relevance. Shark search also somewhat smooths the depth cutoff method that its predecessor used. Meanwhile, Junghoo Cho, Hector Garcia-Molina, and Lawrence Page have proposed reordering the crawl frontier according to page importance, 3 which they can compute using various heuristics—page rank, number of pages pointing to a page (in-links), and so on. These three algorithms don't employ a classifier, but rather rely on techniques based on information retrieval (IR) to determine relevance.
Modifies fish-search in two ways: child inherits discounted value of ancestor; max over ancestors Consider anchor and context of anchor as well as text.

4.6 Tunneling

Sometimes, pages of the same topic do not point directly one anther and therefore it is necessary to go through several off-topic pages to get to the next related one. Bergmark et al. [9] suggest allowing the crawl to follow a limited number of bad pages in order to reach the good one, naming this technique is tunneling. An essential weakness of the baseline crawler is its lack of ability to model tunneling that is, it can't tunnel to the on-topic pages by following a path of off- topic pages. Two remarkable projects, the context-graph-based crawler and Cora's focused crawler, achieve tunneling. [12]

4.7 Info-spiders

Collection of agents (spiders). Agent state = vector of keywords
at each interation, a randomly chosen agent:

1. chooses a link, based probabilistically on match of state with context of anchor;
2. retrieves document D;
3. Gains energy sim(Q,D) - Cost; (Can be negative).
4. Modifies state vector using new document, neural net (or perception).
5. If energy great enough, produces offspring with mutated state; divides energy with offspring.
6. If energy too low, dies.

Note: An *agent* learn clues from anchor, which it passes to offspring. [12]

4.8 Genetic Algorithm

Genetic algorithms (GA) are search algorithms based on the principle of natural selection and genetics. GA operates on a population of potential solutions applying the principle of the survival of the fittest to produce better and better approximation to the solution of the problem that GA is trying to solve. At each generation, a new set of approximations is created by the process of selecting individuals according to their level of fitness value in the problem domain and breeding them together using the operators (crossover and mutation) borrowed from the genetic process performed in the nature. This process leads to the evolution of populations of individuals that are better adapted to their environment than the individuals that they were created from, just as it occurs in natural adaptation. [5]

5.CONCLUSION

The main aim of the review paper was to throw some light on the web crawling algorithms.

Due to the dynamism of the Web crawling forms the back-bone of applications that facilitate Web information retrieval. While the usual use of crawlers has been for creating and maintaining indexes for general purpose search- engine. A number of topical crawling algorithms have been proposed in the literature. Building an effective web crawler to solve the purpose is not a difficult task, but choosing the right strategies and building an effective architecture will lead to implementation of highly intelligent web crawler application.Genetic Algorithm due to its iterative selection from the population to produce relevant results. The focused crawler is a system that learns the specialization.

REFERENCES:

[1] Michelangelo Diligenti, Frans Coetzee, Steve Lawrence, C. Lee Giles, Marco Gori, " Focused Crawling using Context Graphs", 26th International Conference on Very Large Databases, P. 527 - 534 2000.

[2] S. Chakrabarti, M. van der Berg, and B. Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," in Proc. 8th International World-Wide Web Conference toronto, p. 545-562, 1999.

[3] J. Cho, H. Garcia-Molina, and L. Page, "Efficient Crawling Through URL Ordering," Proceedings of the Seventh International World Wide Web Conference. Netherlands, Volume 30, April, P. 161-172, 1998.

[4] S. Chakrabarti, M. van der Berg, and B. Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," in Proc. 8th International World-Wide Web Conference toronto, p. 545-562, 1999.

[5] C.Singh, Improving Focused Crawling With Genetic Algorithms. International Journal of Computer Applications 66(4):40-43, March 2013. Published by Foundation of Computer Science, New York, USA ISBN: 973-93-80873-64-0

[6] Mohsen Jamali et. al "A Method for Focused Crawling Using Combination of Link Structure and Content Similarity" Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence,2006.

[7] Zhaoqiong Gao et. al," Incrementally Updating Concept Context Graph (CCG) for Focused Web Crawling Based on FCA", in proc. Asia-Pacific Conference on Information Processing, vol. 2, p.40-43, 2009.

[8] C. Singh "Domain Specific Collection With Genetic Algorithms" published in Dronacharya research journal Vol. –IV Issue-1 Jan-june-12 ISSN 09753389

[9] Milad shokouhi, Pirooz Chubak, Zaynab Raeesy," Enhancing Focused Crawling with Genetic Algorithms," Information Technology: Coding and Computing, Volume 2, Issue, 4-6 April P. 503 – 508, 2005.

[10] Knut magne risvik and Rolf michelsen, "Search Engines and Web Dynamics" Computer Networks Volume 39, Issue 3, 21 June, P. 289-302, 2002.

[11] Marc najork and Janet wiener "Breadth-First Search Crawling Yields High-Quality Pages" WWW10, May 1-5, Hong Kong, 2001

[12] C. Singh, Ramkala. Article: "Web Crawling Algorithms" page No. 161-165, ID-raictia-10 - 194.

[13] Alessandro Micarelli and Fabio Gasparetti,"Adaptive Focused Crawling" Springer-Verlag Berlin LNCS 4321, Heidelberg p.231–262, 2007

[14] MPS Bhatia, Akshi Kumar Khalid, "A Primer on the Web Information Retrieval Paradigm" Journal of Theoretical and Applied Information Technology, Accepted June 24, p. 657-662, 2008.

[15] Gautam Pant, Padmini Srinivasan1, and Filippo Menczer, "Crawling the Web" Web Dynamics: Adapting to Change in Content, Size, Topology and Use, Springer-Verlag, Berlin, Germany, p.153-178, November 2004.

[16] Anshika Pal, Deepak Singh Tomar, S.C. Shrivastava, "Effective Focused Crawling Based On Content And Link Structure Analysis" International Journal of Computer Science and Information Security, Vol. 2, No. 1, June 2009.

[17] Qu Cheng et. al. "Efficient Focused Crawling Strategy Using Combination of Link Structure and Content Similarity" Proceedings of IEEE International Symposium on IT in Medicine and Education. vol.2, July, p.797 – 802, 2003.

[18] Shalin Shah "Implementing an Effective Web Crawler".