Focused Web Crawler for learning contents

Rajat Kumar Sharma [1], Deepak Kumar[2]

[1]*MCA(LE)4thsem, CCSIT, TMU ,Moradabad*
[2]*Assistant Professor, CCSIT, TMU, Moradabad*
[1]*rajatraj471135@gmail.com*
[2] deepakchaudhary008@gmail.com

Abstract—**Anhuge amount of learning stuff is needed for the e-learning content management system to be useful. This has lead to the difficulty of locating proper learning stuff for a particular learning topic, creating the need for automatic searching of good content within the learning context. In this paper, our aim to deal with this need by proposing a novel approach to find out good materials from www for eLearning content management system. This work presents domain ontology concepts based query method for searching documents from web and proposes concept and term based ranking system for obtaining the ranked seed documents which is then used by a concept-focused crawling system. The set of crawled papers so obtained would be obtained anproper set of content matter for building an e-learning comfortablemanagingmethod.**

Keywords—***Content search and retrieval, domain ontology, ranking system, concept-focused crawling, eLearning.***

I. Introduction

Crawlers can validate hyperlinks and HTML code. They can also be used for web scraping called a Web spider, an automatic indexer or Web scatters. A Web crawler starts with a list of URLs to visit, called the seeds. As the crawler visit these URLs, it identifies all the hyperlinks in the page and adds them to the record of URLs to visit, called the crawl boundary. URLs from the frontier are recursively visited according to a set of policies. Various techniques used for web crawling are also explained to get the knowledge for focused web crawling. The www is a combination of billions of related documents formatted using HTML. In a Web, a user views the Web pages that contains images, text and other multimedia and navigates between them using hyperlinks. When you ask a search engine to get the desired information, actually searches through the index which it has created and does not actually searches through the Web. The information can be used to gather more on associated data by cleverly and professionally choosing what links to follow and what pages to discard. This process is called Focused Crawling.[1]

II. Focused Web Crawling

Focused Web crawling is integrated piece of infrastructure for search engines. Knowledge workers are available more in India. Indians are working in all over the world in Knowledge seeking Industries. In Indians day-to-day life, they seek for a search engine to gather knowledge. Search Engines are the main source of information used by Indians to gather knowledge. Search engines are intended to help users find relevant information on the Internet. Typically users summit a query to a search engine, which returns a list of links to pages that are most relevant to the query. the web is changing all the time due to the appearance of new documents and the disappearance or changing of old ones, crawlers need to work incrementally and continuously avoiding repeating jobs[2]. A Web crawler starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier. URLs from the frontier are recursively visited according to a set of policies.

III. Crawling Policies

The Web crawler is the outcome of a combination of policies:

A. Selection Policy

That states which pages to download, even hugeexplore engines cover only a part of the widelyexisting part.

B. Re-Visit Policy

That states when to check for changes to the pages, By the time a Web crawler has finished its crawl, many events could have happened, including creations, updates and deletions.

C. Politeness Policy

That states how to avoid overloading sites. If a single crawler is performing multiple requests per second, a server would have a hard time keeping up with requests from multiple crawlers.

D. Parallelization Policy

A parallel crawler is a crawler that runs several processes in parallel. The aim is to make the most of the download speed while minimizing the transparency from parallelization and to avoid duplicate download.

IV. CRAWLING PROCESS

A Search Engine Spider is a program that most search engines use to find what's new on the Internet. Google's web crawler is known as Google bot. There are many types of web spiders in use, but for now, we are only interested in the Bot that actually —crawls‖ the web and collects documents to build a searchable index for the different search engines. The program starts at a website and chases all hyperlink on each page. So we can say that everything on the web will finally be start and spider, as the so called —spider‖ crawls from one website to another. Search engines may run thousands of instances of their web crawling programs concurrently, on several servers. When a web crawler visits one of your pages, it loads the site's content into a database. Once a page has been fetched, the text of your page is loaded into the search engine's directory, which is a huge database of words, and where they happen on different web pages. Basically three steps that are involved in the web crawling process:-

• The search crawler starts by crawling the pages of your site.

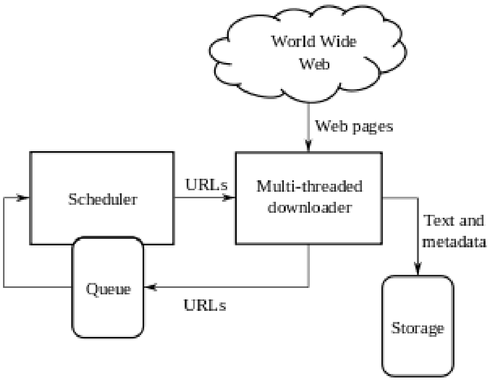
• Then it continues indexing the words and content of the site.

• Finally it visits the links (web page addresses or URLs) that are found in your site.

When the spider doesn't find a page, it will finally be deleted from the directory. However, some of the spiders will check again for a second time to verify that the page really is offline. The first thing a spider is supposed to do when it visits your website is look for a file called —robots.txt‖. This file contains instructions for the spider on which parts of the website to index, and which parts to ignore. The only method to manage what a spider sees on your site is by using a robots.txt file. All spiders are supposed to follow some rules, and the major search engines do follow these rules for the most part. Providentially, the main search engines like Google or Bing are finally working jointly on standards. When the Google spider looked at an HTML page, it took note of two things:

• The words within the page

• Where the words were found



V. SOME CLASSIFICATION APPROACHES

Helpfulapplicationcalculation can help pass up downloading and visiting variousunrelated pages. New learning-based approach which uses the Naïve Bayes classifier as the base prediction model to improve relevance prediction in focused Web crawlers is used. A learning-based focused crawler has learning skill to adjust to its search topic and to progress the exactness of its prediction of relevancy of unvisited URLs. New learning-based focused crawling approach that uses four relevance attributes to predict the relevance and if additionally allowing dynamic update of the training dataset, the prediction accuracy is further boosted. [3]

While surfing the internet it is hard to deal with unrelated pages and to expect which links lead to excellence pages. In this paper, a technique of effective focused crawling is implemented to improve the quality of web navigation. To check the similarity ofweb pages w.r.t. topic keywords, a similarity function is used and the priorities of focused crawler. In this approach, URL queue contains a list of unvisited URLs maintained by the crawler and is initialized with seed URLs. Web page downloader fetches URLs from URL queue and downloads equivalent pages from the internet. The parser and extractor extract information such as the terms and the hyperlink URLs from a downloaded page. [4]

Relevance calculator calculates relevance of a page w.r.t. topic, and assigns make to URLs take out from the page. Topic filter analyzes whether the content of parsed pages is related to topic or not. If the page is related, the URLs extract from it will be additional to the URL queue, otherwise added to the unrelated table and this approach has better presentation than the BFS crawler. [5]

Intelligent crawling method involves looking for specific features in a page to rank the applicant links. These features include page content, URL names of Web page, and the personality of the parent and sibling pages. It is a generic framework in that it allows the user to specify the relevant criterion. The system has the ability of self-learning, i.e. to collect statistical information during the crawl and adjust the weight of these features to capture the dominant individual factor at that moment. [6]

The Fish-Search is an untimely crawler that prioritizes unvisited URLs on a queue for definite search aim. The Fish-Search approach assigns priority values (1 or 0) to candidate pages using simple keyword matching. One of the disadvantages of Fish-Search is that all applicable pages are assigned the same priority value 1 based on keyword matching. Fish-Search the Web is crawled by a team of crawlers, which are viewed as a school of fish. If the fish finds aapplicable page based on keywords specified in the query, it continues looking by following more links from that page. If the page is not applicable, its child links receive a low privilegedvalue.[7]

The Shark-Search is a modified description of Fish-Search, in which, Vector Space Model (VSM) is used, and the priority values (more than just 1 and 0) are computed based on the priority values of parent pages, page comfortable, and anchor text. The Shark-Search is modification of Fish-search which differs in two ways: a child take overs a discounted value of the score of its parent, and this score is combined with a worth based on the anchor text that occurs approximately the link in the Web page. Extracted out links are also calculated based on Meta data and resultant pages generated fromA relatively more recent type of focused crawlers adopts learning-based approaches to relevance guess. Info Spiders and Best-First are the examples of focused crawling methods. [8]

Best-First methods apply VSM to work out the significance between candidate pages and the search topic. Shark-Search crawlers may be judged as a type of Best-First crawlers, but the former has a more confused function for computing the priority importance. In, Best-First was shown most winning due to its cleanness and effectiveness. *N*-Best-First is generalized from Best-First, in which *N* best pages are chosen as a substitute of one. [9]

The HITS algorithm is another method for rating the quality of a page. It introduces the idea of authorities and hubs. An authority is a prominent page on a topic. They are the target of the crawling process since they have high quality on a topic. A hub is a page that points to many a characteristic is that it's out links are suggestive of high quality pages. Hubs do not need to have high quality on the topic themselves or links from good pages pointing to them. The idea of the hub is a solution to the problem of distinguishing the popular pages, from the authoritative pages. Therefore, hubs and authorities are defined in terms of mutual recursion.

Depth-first crawling trails everyprobable path to its ending before another lane is tried. It works by finding the first link on the first page. It then crawls the page link with that link, finding the primary link on the new page, and so on, in anticipation of the end of the path has been reached. The process continues until all the branches of all the links have been beat.[10]

Ontology based focused crawling utilizes the notion of ontologies in the process of crawling. It consists of two main processes which interact with each other, the ontology cycle and the crawling cycle. In the ontology cycle, the crawling aim is defined by ontologies (provided by the user) and the articles that are considered related as well as suggestions for the improvement 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]

Reinforcement Learning (RL) Crawler is used to train a crawler on specified example web sites containing target documents. The web site or server on which the document appears is repeatedly crawled to learn how to construct optimized paths to the target documents. [12]

Neural Networks extends the RL method for focused crawling. In their approach, each webpage is represented by a set of 500 binary values, and the state of each page is determined by Temporal Difference Learning, in order to minimize the state space. The relevance of the page depends on the presence of a set of keywords within the page. A neural network is used for the estimation of the values of the different stages. [13]

The work is most related to. In, a classical focused crawler was proposed. Several improvements in terms of speed and prediction accuracy were made in and resulted in a learning-based crawler. In [13], out-links carry relevance scores from the originating/parent pages.

However, unrelated pages are not overlooked immediately, but broken for discovering new related pages led from these unrelated pages. This mechanism will be discussed and used in implementation. In this two more classifiers are added which are Neural Network and Decision Tree Induction .The relevance characteristics, i.e., parent pages and contiguous texts, attach text and URL word relevancy.

VI. Conclusions

Focused crawler is a web crawler that collects Web pages that satisfy some specific property, by carefully prioritizing the crawl frontier and managing the hyperlink exploration process. Some predicates may be based on simple, deterministic and surface properties.

A. From the results it is clear that the features detected i.e. URL text relevancy, Surrounding text relevancy, Parent Pages relevancy and Anchor text relevancy were decent enough for the vigorous feature detection that are intended to improve.

B. Work was to compare the precision rate of three prime classifiers.

C. The analysis shows that in terms of classification precision rate, time and complexity neural network leads the Decision tree induction and Naive Bayesian by a big margin.

D. Moreover in order to have more complex computing need to improve upon the memory of a particular classifier by training it, in this context also the NN dominance prevails over the DTI and NB.

REFERENCES

[1] Chakrabarti, S., van den Berg, M. & Dom, B., 1999a Focused crawling: a new approach to topic-specific Web resource discovery. In Proceeding of the 8th International conference on World Wide Web,Canada, pp.1623

[2] J. Rennie and A. McCallum, "Using Reinforcement Learning to Spider the Web Efficiently," In proceedings of the 16th International Conference on Machine Learning(ICML-99), vol.14,pp. 335-343, 1999.

[3] Mejdl S. Safran1,2, Abdullah Althagafi1 and Dunren , ‖ Improving Relevance Prediction for Focused Web Crawlers ‖ , IEEE/ACIS 11th International Conference on Computer and Information Science,2012

[4] D. Taylan, M. Poyraz, S. Akyokus, and M. C. Ganiz, ― Intelligent Focused Crawler: Learning Which Links to Crawl ‖ , In Proc. Intelligent Systems and Applications Conference (INISTA) , pp. 504-508, 2011.

[5] Kleinberg, M. J.,. ‖ Authoritative Sources in a Hyperlinked Environment, ‖ In Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithm,1997 Global Journal of Computers & Technology Vol. 3, No. 2, October 06, 2015 www.gpcpublishing.com ISSN: 2394-501X 172 | Pageeditor@gpcpublishing.com

[6] Y. Guo, K. Li, K. Zhang, and G. Zhang, ‖ Board forum crawling: a Web crawling method for Web forum ‖ , In IEEE/WIC/ACM Int. Conf. Web Intelligence, Hong Kong,vol.4,pp 745-748,2006.

[7] Christopher Olston and Marc Najork, ―Web Crawling ‖ ‖ , Foundations and Trends in Information Retrieval, Vol 4.No 3(2010) C.

[8] S. Chakrabarti, M. Berg, and B. Dom, ―Focused Crawling: A New Approach for Topic Specific Resource Discovery ‖ , In Journal of Computer and Information Science, vol. 31, no. 11-16, pp. 1623-1640,1999.

[9] Li, J., Furuse, K. & Yamaguchi, K., ― Focused Crawling by Exploiting Anchor Text Using Decision Tree ‖ ,In Special interest tracks and posters of the 14th international conference on World Wide Web,Chiba, Japan,2005.

[10] Mejdl S. Safran1,2, Abdullah Althagafi1 and Dunren , ‖ Improving Relevance Prediction for Focused Web Crawlers ‖ , IEEE/ACIS 11th International Conference on Computer and Information Science,2012

[11] D. Taylan, M. Poyraz, S. Akyokus, and M. C. Ganiz, ― Intelligent Focused Crawler: Learning Which Links to Crawl ‖ , In Proc. Intelligent Systems and Applications Conference (INISTA) , pp. 504-508, 2011.

[12] Aggarwal, C., Al-Garawi, F. & Yu, ― 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 – 105,2001

[13] De Bra, P. and Post, R., ‖ Information Retrieval in the World_Wide Web: Making Client_based searching feasible ‖ , In Journal on Computer Networks and ISDN Systems, pp. 183-192, 199