Effective implementation of focused web crawler using scalable and extensible architecture

Dr. P. Jaganathan [1], T. Karthikeyan [2]

[1.] Professor and Head, Department of Computer Application, PSNA College of Engineering and Technology, Dindigul, India
[2.] Research Scholar, Department of Research and Development Centre, Bharathiar University, Coimbatore, India
karthik.rt@gmail.com

Abstract: The rapid growth of the World Wide Web contents makes the focused crawler a challenging task. To selectively download relevant pages from the World Wide Web is the problem that focused crawlers deal with it. In this paper we present a Scalable and Extensible Web Crawler with Focused Web Crawler, this makes the major components of any Scalable and Focused Web Crawler then describe the particular components used in this Architecture. It also support for users to download support information and Extensibility. This method also describe how the focused Web crawler components and A Scalable, Extensible Web Crawler was integrated. It also describes how to work together and their functionality of every component. In minimum time this Architecture downloaded maximum pages from web and extract web pages partially which is needed to users. This focused crawler cannot be categorized to the existing focused crawlers approach since it has their own features.

[T. Karthikeyan, P. Jaganathan. **Implementation of effective focused web crawler using scalable and extensible architecture.** *Life Sci J* 2013;10(3):1967-1974] (ISSN: 1097-8135). http://www.lifesciencesite.com. 291

Keywords: focused crawler, Ontology-based focused crawlers, Structure base focused crawlers, User's query.

1. Introduction

In the usage of internet the advent of the World Wide Web has caused a dramatic increase. A wide range of information can be obtained at a low cost in a broadcast medium called World Wide Web. Information on the WWW is important to the individual users and also to the business organizations especially when concerning about the critical decision-making.by using a combination of search engines and browser most users obtain WWW information, however these two retrieval mechanisms do not address all of a user's information needs. It's true in the business Organizations that to analyze these data to discover useful knowledge to support decision making and to systematically harness strategic information from the Web.

Typically for the purpose of Web indexing A Web crawler is an Internet based that systematically browses the World Wide Web .it may also may called as a Web spider , an automatic indexer, an ant or a Web scutter. Over Million Pages having in World Wide Web and it continues to grow rapidly which makes it difficult, to exact user's needed information. This proposed Architecture has a Scalable, Extensible Web Crawler using focused web crawler. General purpose search engines retrieve and download unwanted information therefore load gets increased in the network .so this proposed Architecture first take users query , then important word is extracted and eliminates bibliography word, then check in database, if the word not found in database then find similar keywords from world dictionary and make a world database. The word database connect to the focused web crawler and their URLs .this URLs store in URL database and connect to Scalable , Extensible Web Crawler, then this web Crawler download the exact information as of user's needs.

General crawlers judge whether the documents pointed by the URLs are related to the specific domain, but the focused crawlers vary from that. Based on the relevance probabilities ordering the URL queue and the pages are very relevant for the specific domain which are downloaded first. Focused crawlers selectively look for the relevant web pages and ignore the irrelevant ones. The benefit of the focused crawling approach on that specific domain is able to find relevant documents of large numbers and is able to discard irrelevant documents, which leads to significant savings in both communication and computation resources and high quality retrieval results.

2. Material and Methods

2.1 literature survey

Since the search engine have come into existence, development of various techniques were witnessed in the literature, in order to get optimized result of the search engine, some of the techniques were focused on the ranking of the search results while others are related to the crawling appropriate pages depending upon end user search engine. In order to achieve these objectives both statistical and natural language is processed. For example a dependency analysis based link- context was extracted. The mail idea is to simulate the browsing behavior of web readers. The author fractionize the behavior into four steps which were passing, decomposing, grouping and selection. but in this technique as author itself made a statement that word variation between the link- context and the target web-page has made the quality of link-context derivation very low. Authors have described an approach to generate automatic rich semantic annotations of text which can be utilized by semantic-web. Authors have given an idea that cohesive text and non-cohesive text surrounding. It provides rich semantic cuse about a target web-page. In a scheme based, on passing of the text around anchor-link was proposed.

2.2 Focused crawler

The focused crawler, during which a crawler seeks, acquires, indexes, and maintains pages on a particular set of topics that represent a comparatively slim phase of the net. The focused internet crawlers square measure designed for retrieving web content supported the principles that determine relevant pages or/and priority criterions to sequence the net pages to be crawled and add them to the native information . Focused crawlers aren't designed just for downloading documents to be indexed for a website specific program. However they are also designed to downloaded documents to use as a supply for data processing. Focused travel could be a hopeful approach for rising the preciseness and recall of professional search on the net. As aforementioned before, the focused travel motivation comes from the poor performance of general purpose search engines, those rely upon the results of generic internet crawlers. The focused crawlers aim to look and retrieve web content from the globe Wide internet, that square measure associated with a particular domain. Rather than visiting all web content, a focused crawler visits solely the region of the net that contains relevant pages, attempting to skip moot regions. This results in vital savings in each computation and communication resources.

The method focused crawlers exploit hyper-textual data is one in all the options that characterize them. Ancient crawlers convert an internet page into plain text extracting the contained links, which can be accustomed crawl alternative pages. Focused crawlers exploit extra data from web content, like anchors or text close the links. Essential problems with focused crawler is the way to determine links and pages that square measure relevant to the precise domain, and to order the universal resource locators within the URL queue . So, a focused crawler needs to predict exactly an internet page's connection before downloading it. Early focused crawlers relay on victimization the domain keywords to see if the page has relevancy or not once downloading it. Attempting to boost the focused crawler, some use metaphysics to notice the relevant score for links before downloading. They order the antecedent transfer documents victimization metaphysics by computing the page connection. The downloaded document continues to be used. At that time they filter the retrieved documents by applying SVM classification. Some researches analysis the structure hyperlinks to judge the connection. A focused crawler approach, which evaluates the page's content connection victimization metaphysics and hyperlinks analysis. Victimization the link structure analysis with the similarity of the page context to see the transfer pages priority. We will categorise the focused crawler approaches in line with their dependency on determinant the relevant pages to: metaphysics based mostly focused crawler, structure based mostly focused crawler and focused crawler approaches.

3. Discussion

3.1 Ontology primarily based focused crawler

Generally ontology primarily based focused crawlers square measure series of crawlers, to link the fetched net documents with the domain ideas they use ontology's. Ontology base focused crawlers relay on utilizing the domain ideas to gauge the page connection. Most of researches use ontology to gauge the connection before downloading the net pages. Authors verify the connection score for the new transfer pages consistent with the connection of the net contain of the page that has coupled to their uniform resource locator. Whereas others use the words as regards to the links to calculate the connection of the coupled uniform resource locator. Others use ontology to filter the retrieved pages. Crawler exploits the Web's link structure to retrieve pages by traversing links from antecedently retrieve ones. As pages square measure fetched, their outward links could also be extra to an inventory of unvisited pages, that is concerning because the crawl frontier. To spot consequent most acceptable link to follow from the frontier, they used AN ontology primarily based rule to cipher page connection.

The ontology ideas square measure extracted from the transfer page and square measure accustomed calculate the connection of the page. Then a candidate list of sites so as of accelerating priority is maintained. supported the page ranking a call are going to be taken to feature its links to be crawl or no. they need AN ontology learning module in their approached use the transfer pages to bootstrap their ontology. An attempt to improve the present focused crawl by exploitation the construct with its context and context data for retrieving net documents. Their focused crawler starts by crawl a given set of URLs. This connection live may be a operate that tries to map the content of an internet document and its existing, already collected context data to realize AN overall connection score. The given score depends on the amount of correlation. Sentence question similarity is employed to see the connection.

A priority queue is maintained so relevant documents square measure downloaded first. CORE may be a focused crawler design that retrieves the documents primarily based upon their ontology connection score. It additionally takes care of the tunneling (the cases wherever one will reach to the relevant pages through some extraneous pages). It starts by seed of words that square measure taken from user. CORE uses these seeds to retrieve their uniform resource locator from the gawk, Yahoo and MSN search engines. The seeds uniform resource locator square measure prioritized in 3 classes: High, Medium, and Low. Crawler manager downloads the document consistent with the uniform resource locator priority queue and store it within the document repository. New URLs square measure extracted from documents within the Document Repository and therefore the URLs with the encircling text that embrace a hard and fast variety of letters proceeding and succeeding the link, the heading or sub heading beneath that the link seems is extracted from the document. The extracted link with the context data is passed to the ontology primarily based relevant score that calculate the connection score for every context link. Luong and his faculties retrieve documents and knowledge in an exceedingly biological domain by exploitation search engines and digital libraries. The retrieved documents square measure mechanically separated out the documents that square measure really relevant to the biological domain of interest. They mechanically generate queries for every construct within the hierarchically. These queries square measure submitted to a range of net search engines and digital libraries. They apply SVM classification to separate out retrieved documents that match the question well however that square measure less relevant to the domain of amphibian morphology ontology.

A set of documents relevant to amphibian morphology square measure created. Their system starts by a manual ontology that is enrich by mining the knowledge extracting from the crawled documents. Principle and Hsu gift an ontology-support net focused-crawler (Onto Crawler) wherever the user entered some keywords. These keywords weren't completed and not capable to obviously indicate the question demands of users. Thus their system use the domain ontology to supply comparison and verification for those keywords thus on up-rise the preciseness and recall rates of webpage looking out. Their crawler starts by causing question to Google and Yahoo question uniform resource locator, that come back the question results from search engines, that typically contains an inventory of URLs. The matched ones will be downloaded. Reading the content of the link file to evaluate whether or not the webpage was placed within varies of the question or not. They compare the content of the webpage with the ontology to create this decide.

3.2 Structure primarily based focused crawler

Structure base focused crawlers soak up accounting the net pages structure once evaluating the page connectedness. The researchers analysis the hyperlinks between the candidate crawled page and also the domain sites and to work out if it relevant to the domain or not. Others use all hypertext markup language components to work out the relevant of the net pages .All others plan is predicated on it, the standard hyperlinks in pages square measure a illustration to the author's read regarding alternative pages. Additionally the contents of pages square measure another supply to relate them to a website. Their crawler starts with one seed page and tries to fetch the foremost connected pages to specific domain (Sports) from the net. The links during this page square measure thought of as an initial seed pages to be downloaded. The links pages in every seed square measure transferred and keep during a download folder. for every transfer page work out and store its similarity degree to the domain, variety of links from the page to seed pages and variety of links from seed pages thereto. They use the keywords and phrases utilized in the domain to calculate the similarity.

The candidate crawled pages square measure ordered supported a mixture of those 3 values keep for every page. The page with the very best rank is chosen and other to seed pages. Huang gift AN intelligent focused crawler approach that evaluates the page's connectedness to a selected domain by victimization each the domain metaphysics and also the hyperlinks association to the domain sites. They transfer sites and break down their links and content info. When preprocessing of the visited sites, the entities (relevant metaphysics concepts) and also the links‟ structures square measure extracted and counted. The links square measure analyzed by obtaining the all link info and formulates the link priority score by the predictor. The content entities square measure accustomed calculate the content connectedness score by the entities mapping to the E-commerce metaphysics. The score of crawl priority is predicated on the link and content priority score arranges to the sequence of travel in line with the crawl priority score. The larger crawl score pages are visited a lot of quickly. Their approach has the aptitude to bootstrap the used metaphysics by employing a machine learning algorithms.

A link relates the relevant page with an anchor occurring among a downloaded page, and also the anchor itself has any structural relationships with alternative hypertext markup language components in its link context, etc. They discover this relation by learning classification rules supported first-order logic. First, they tagged instances with their data|background|information} (using the knowledge of the link structure of the pages) to represent the connectedness clues of the pages as facts. Then a relative learning algorithmic program uses these facts to find a group of first-order rules. The discovered rules square measure accustomed guide the focused travel method.

They use the text of sure hypertext markup language components to priorities documents for downloading. And so considerably improve the speed of convergence to a subject. They study the result of structured connectedness analysis on focused travel. Their study showed that the power to extract anchor text from hypertext markup language documents permits their focused crawler to extend the speed of initial topic acquisition. The crawler has additionally shown higher performance in maintaining an antecedently no inheritable topic-specific assortment.

4. Architecture for Scalable, Extensible Web Crawler

A Scalable, Extensible Web crawler which is parallel down load pages like .pdf, .text, .doc, .html, jpeg, etc. These files are downloaded in parallel.

Scalable

To scale up to entire web This Web Crawler Architecture is designed, to fetch of millions of web documents it has been used. The vast majority of our data structures and small parts are stored on disk and memory for efficiency.

Extensible

In modular way This Web Crawler Architecture is designed, that new functionality will be added by third parties with the expectation.

Proposed System Architecture

The proposed Architecture connected with world dictionary for finding out similar meaningful world and in word database these word are stored , then this word's database word one by one access by Focused Crawler, then create URL database .And these URL documents downloaded by A Scalable , Extensible Web Crawler as shown in figure 2.

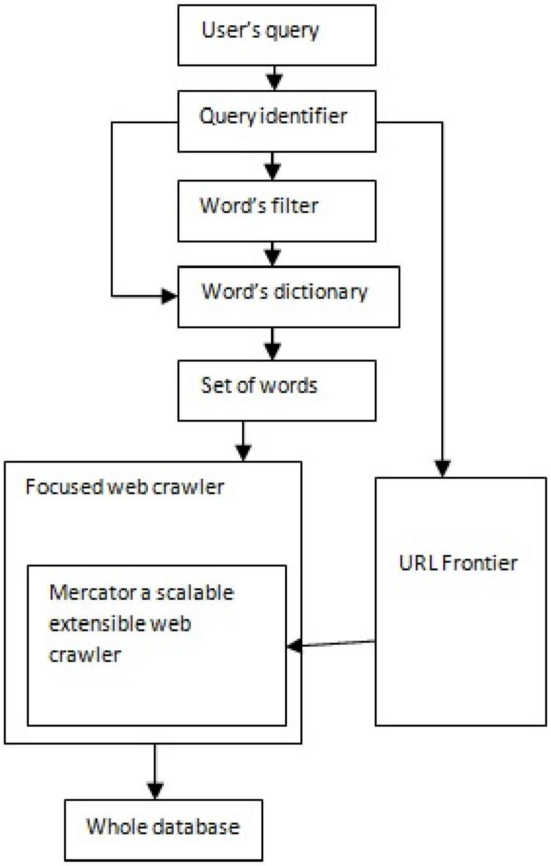


Figure 1. System architecture flow

Problem identification

In the Existing System it download relevant pages which is related to the user's question and Focused Crawler can download more relevant pages which is related closely to the user's query. Normal Web Crawler can't download more relevant pages and Focused Web Crawler can't download .pdf files, .doc files, .text files etc in parallel. Our Architecture Scalable, Extensible Web Crawler with Focused Web Crawler can Download .pdf files,. Text files, .doc files, .html files, .xml files in parallel and related closely to user's query.

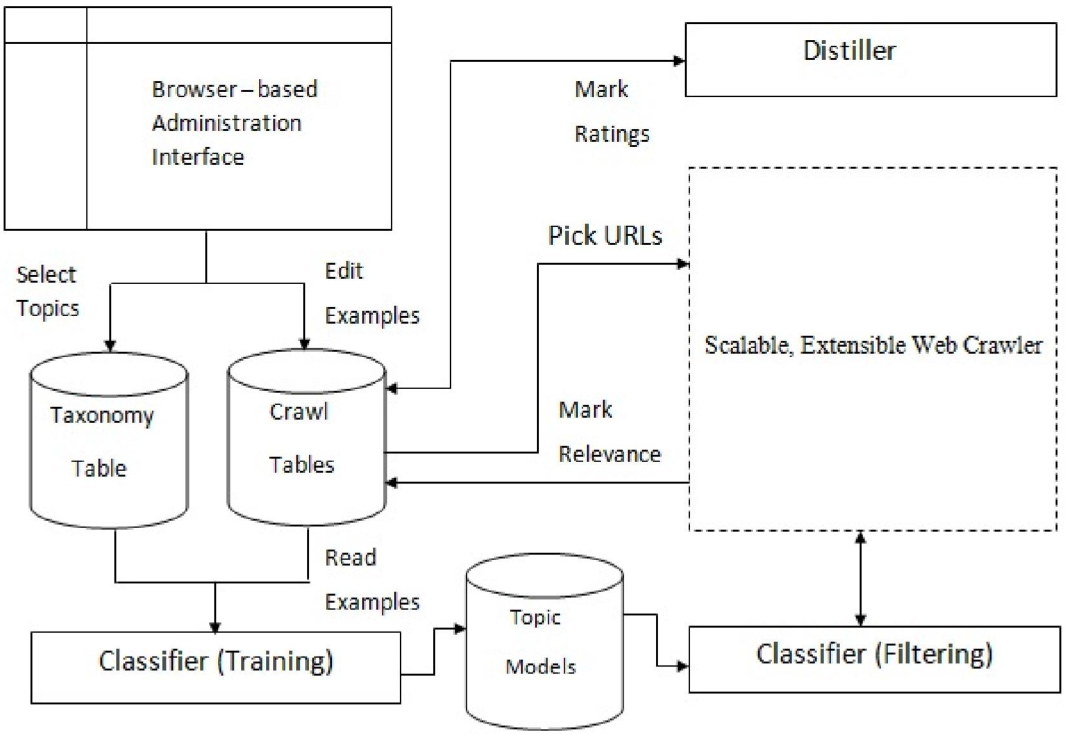


Figure 2. Proposed architecture: A Scalable, Extensible web Crawler, Classifier, Distiller are integrated

5. Implementation and result

The collection of computer addresses could be a URL frontier that the crawler intends to fetch and method within the future. Computer address address frontiers typically add one among 2 ways that - a batch crawl or a nonstop crawl. A batch crawl's frontier contains solely new URLs, whereas a nonstop crawl's will contain a seed set of URLs however new ones is also more (or existing ones removed) throughout the crawl.

Regardless of the sort of crawl, the information store accustomed manage the frontier must be quick as a result of it'll weigh down creep otherwise. Figure 3 that helps illustrate the URLs in an exceedingly crawl frontier.

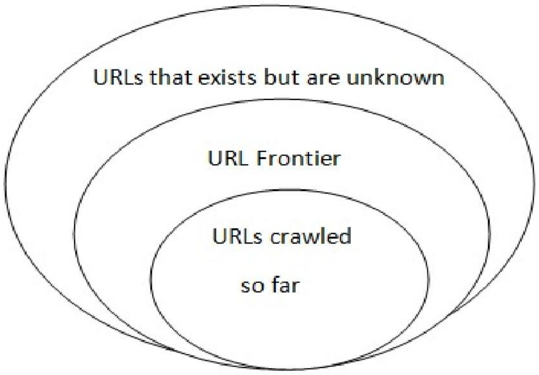
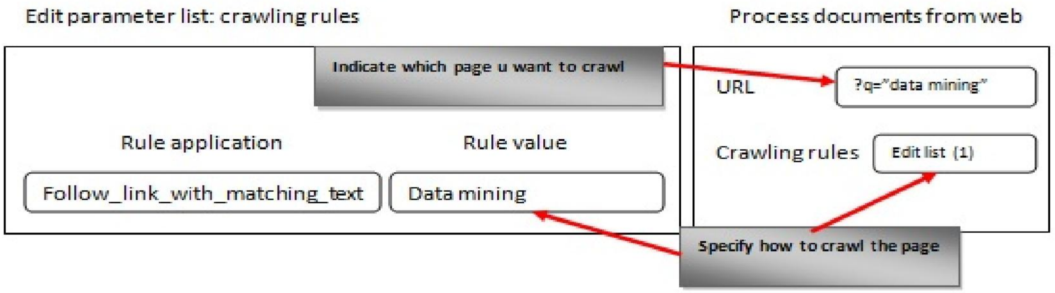


Figure 3. URLs in an exceedingly crawl frontier


Edit parameter list: crawling rules
Process documents from web
Indicate which page u want to crawl
Specify how to crawl the page
Edit list (1)

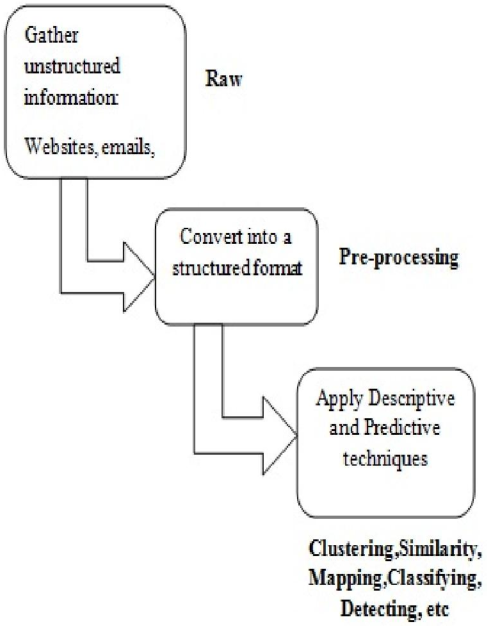
URL ?q="data mining"

Crawling rules

Rule application Follow_link_with_matching_text

Rule value Data mining

Figure 4. Initiating web crawl on a specified URL


Gather unstructured information:
Websites, emails,
Raw
Convert into a structured format
Pre-processing
Apply Descriptive and Predictive techniques
Clustering,Similarity, Mapping,Classifying, Detecting, etc

Crawling is quite simple at its core:

1. To crawl select a URL
2. Fetch and parse page
3. Saving important content
4. From page Extract URLs
5. Add URLs to queue
6. Repeat

The small print around URL discovery is sort of sophisticated betting on your application .Adding logic around that URLs to pick out to crawl (selection), and that URLs to filter (often times a Simple restriction is to line your filters to solely apply to specific domains and ignore all others) will facilitate make sure that your crawl is tightly targeted to your target content/data.

One way to make your universal resource locator frontier is to outline a seed set of URLs or domains and crawl them to extract and see new links. Post process the links to normalize them will facilitate cut back duplication and may be a sensible best apply to follow. Another example of this post process is filtering the links before adding them to the frontier, like proscribing links to specific domain or victimization. A formula to order the URLs in an exceedingly priority that creates sense the appliance and goals of the crawl.

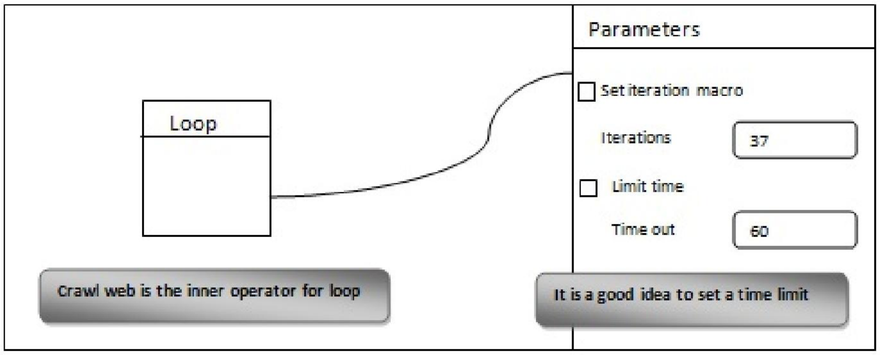
Crawling web

Rapid Miner offers variety of tools to extract info from the net whose reach and power area unit quite spectacular. The fundamental plan is to initiate online crawl on that URL.however we will additionally drill down into the net page by following links.

We glance at 2 totally different operators here. For each operator, you would like to specify creep rules which can permit you to drill down into embedded links inside the page. In this http://www.indeed.com/jobs?q="data+mining" as the url and specify that we need to follow every link in that main url which has the keywords "Data mining" in the anchor text.

Process Documents from Web

It only stores the followed url information unless "add pages as attributes" is checked because this is slightly faster of the two.



Parameters

□ Set iteration macro

Iterations 37

□ Limit time

Time out 60

Crawl web is the inner operator for loop

It is a good idea to set a time limit

Figure 5. Embedding the Crawl Web operator inside the Loop operator

Web Crawl

This enables to store every followed link as a document. To specify the directory into that you want to jot down these files (need to form the directory if it doesn't exist). This is often slightly slower; however it's higher suited to what we have a tendency to do.

Word of caution

Rapid Miner wiki asks you to "please be friendly to the online website house owners and avoid inflicting high traffic on their sites", clearly if you are attempting to crawl too several pages and links, be warned that it the method are going to be slow. Another purpose to notice is in shaping locomotion rules.

Rapid Miner permits you to use any regular expression within the "rule value" - see image higher than. This implies that you simply may have a wild card expression like *Data mining*which is able to pull up any link that has the words in between the * signs. This will additionally slow things down. For our application we'll not use the wild cards however simply persist with the 2 main keywords: information and mining.

Looping

Our objective is to crawl all the work posting that have the word "Data mining" in their anchor text. An easy search on so.com ends up in concerning 400+ jobs announce for this type of position. Clearly this are listed over many pages and that we have to be compelled to traverse through these pages one by one. This is often wherever we'll have to be compelled to use the oop operator. Over 38 pages, 400+ jobs were spread and each page is accessed by a slightly different url as seen below:

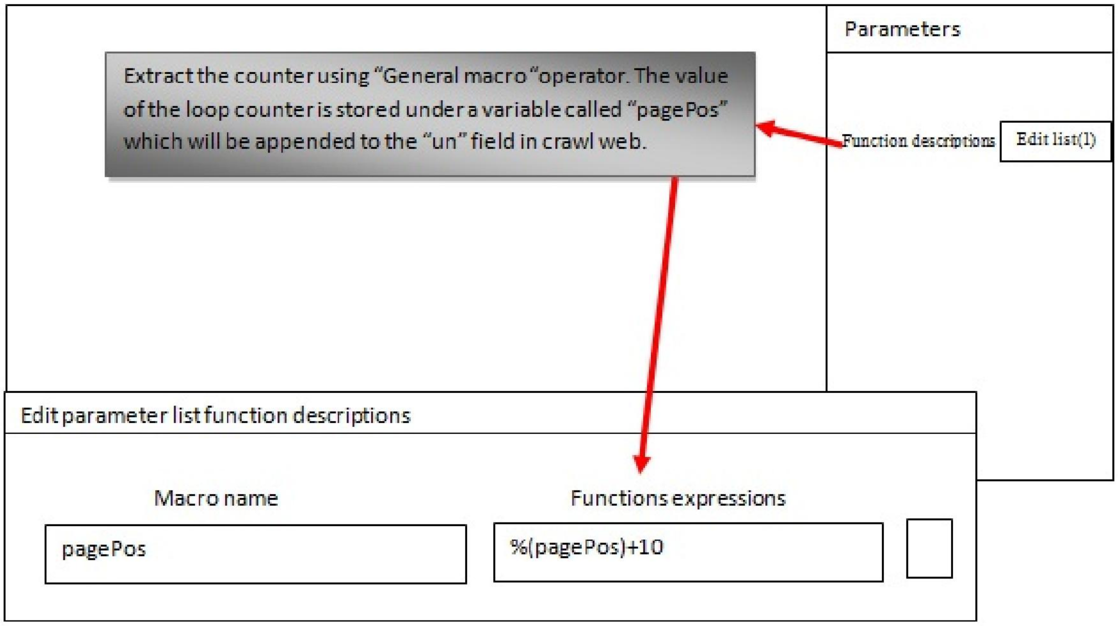
http://www.indeed.com/jobs?q=%22data+scientist%22&start=10

http://www.indeed.com/jobs?q=%22data+scientist%22&start=20

…http://www.indeed.com/jobs?q=%22data+scientist%22&start=370

We need to embed the Crawl Web operator inside the operator Loop and specify 37 iterations for the loop counter

As we have a tendency to bear every iteration, we are going to have to be compelled to update a variable referred to as "pagePos". This is often accomplished by employing a "Generate Macro" operator. In every iteration, we are going to add "10" to the present variable as shown below. You will additionally need a Log operator to stay track.



Extract the counter using "General macro "operator. The value of the loop counter is stored under a variable called "pagePos" which will be appended to the "un" field in crawl web.

Macro name	Functions expressions
pagePos	%(pagePos)+10

Figure 6. "Generate Macro" operator

The pagePos price can then be appended to two fields at intervals the Crawl net operator:

1. The URL field so we tend to advance to following page for crawl.

2. The directory to store the information, whenever we tend to end crawl through one in the entire work posting pages (of the thirty seven during this case), and the crawl net operator can write the output from a link (which could be a hypertext mark-up language page) into a computer file. The names of those files area unit by default zero.txt, 1.txt, and so on. Thus we are going to store every page in its own directory otherwise Rapid Miner will write the saved pages.

6. Conclusion

Thousands of web pages are being added to the web every day. It is becoming difficult to update and crawl the complete web in short time. In such circumstances to a generic crawler a Focused crawling is a promising alternate solution. For a specific domain of user interest Focused crawler aims to search only the relevant subset of the World Wide Web. How to determine the relevant pages is the problem that Focused crawling deals with it. After crawling, the web pages some determine the relevant pages by using domain keywords or ontology. While others before downloading the links they determine the relevant pages. by indexing the web pages Some retrieve focused result using the context associated with the keyword . to determine the relevant pages the pages link structures are used. the social bookmark tags were utilized by others to compute page relevance . by utilizing both the domain ontology and html structure Focused crawlers tried to find relevant web pages before downloading them. Focused crawling approach has a benefit to find a large proportion of relevant documents on a particular domain. it effectively discard irrelevant documents and its able to keep up the change of the Web. The Proposed System Architecture download the user's query which closely relevant documents, text and pdf pages are download in parallel. The Proposed system can Access word dictionary and also search out the synonyms of the words and can download their relevant pages.

Corresponding Author:
Karthikeyan. T
Research Scholar
Department of Research and Development Centre,
Bharathiar University,
Coimbatore - 641 046, India
E-mail: karthik.rt@gmail.com

References

[1] A. Thukral, V. Mendiratta, A. Behl, H. Banati and P. "Bedi, FCHC: A Social Semantic Focused Crawler", in Communications in Computer and Information Science, Vol. 191, Part 5, 2011,pp. 273-283.

[2] M. Kumar and R. Vig, "Design of CORE: context ontology rule enhanced focused web crawler", International Conference on Advances in Computing, Communication and Control (ICAC3"09) pp,2009 494-497.

[3] A. Chandramouli, S. Gauch, and J. Eno, "A Cooperative Approach to Web Crawler URL Ordering", iIn Human Computer Systems Interaction, AISC 98, Part I, pp. , 2012 343–357.

[4] P. Gupta, A. Sharma, J. P. Gupta, and K. Bhatia, „A Novel Framework for Context Based Distributed Focused Crawler (CBDFC)‟, Int. J.CCT, Vol. 1 , No. 1 , 2009, pp.13-26.

[5] H. P. Luong, S. Gauch, and Q. Wang, "Ontology-Based Focused Crawling", Information, Process, and Knowledge Management, 2009 (eKNOW '09) 1-7 Feb. 2009 pp. 123-128

[6] N. Pahal, N. Chauhan, and A.K. Sharma, "Context-Ontology Driven Focused Crawling of Web Documents", A.K. Wireless Communication and Sensor Networks, 2007. WCSN apos: 07. Third International Conference, 13-15 Dec. 2007 pp.121-124

[7] H. Dong, F. K. Hussain, and E. Chang, "State of the art in semantic focused crawlers" in 2009 IEEE International Conference on Industrial Technology (ICIT 2009), Gippsland, in press

[8] M. Bazarganigilani, A. Syed and S. Burki, "Focused web crawling using decay concept and genetic programming", In International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.1, , 2011, pp:1-12.

[9] Abhinna Agarwal, Durgesh Singh, Anubhav Kedia Akash Pandey, Vikas Goel, Design of a Parallel Migrating Web Crawler, International Journal of Advanced Research in Computer Science and Software Engineering , April 2012 Volume 2, Issue 4.

[10] Priyanka-Saxena Department of Computer Science Engineering, Shobhit University, Meerut, Uttar Pradesh-250001, India, Mercator as a web crawler IJCSI International Journal of Computer Science Issues, , January 2012, Vol. 9, Issue 1, No 1

[11] D. J. Abadi, D. Carney, U. C¸ etintemel, M. Cherniack,C. Convey, S. Lee, M. Stonebraker, N. Tatbul, and S. Zdonik. Aurora: a new model and architecture for data stream management. The VLDB Journal, , 2003,12(2):120–139.

[12] A. Heydon and M. Najork. Mercator: A scalable, extensible web crawler. World Wide Web, 2(4): , 1999, 219–229.

[13] P. Jaganthan, T. Karthikeyan and S. Jaiganesh, "A Comparative Study of Various Crawler Algorithms for Web Searching", in International Conference on Mathematical Comp and Management, Jun 17-19, 2010.

7/22/2013