International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 12 | Dec -2016 www.irjet.net p-ISSN: 2395-0072

Efficient and optimized use of data - Shrimp Browser

Ankit Saha, Anurag Singh, Ashok Kumar, Pankaj Kanjani

Student, Department of Computer Engineering, SKN Sinhgad Institute of Technology & Science, Lonavala, SPPU, Pune, Maharashtra, India

---

Abstract - *Web browsing is the process of customizing a Web site to the needs of specific users, taking advantage of the knowledge acquired from the analysis of the user's navigational behavior (usage data) in correlation with other information collected in the Web context, namely, structure, content and user profile data. The primary purpose of a web browser is to bring information resources to the user, allowing them to view the information and then access other information. The normal browsers use large amount of data for uploading and downloading information. Other problem with the normal browsers is, it cannot work if there is no internet connection. Shrimp browser is an application that can work online as well as offline with some advanced features. In online mode users, will be able to get information of n- URL's which are pinned to the browser reducing the time of access, whereas in offline mode some data MB's will be reserved so that user will be updated from time to time even if there is no data connection. We take a list of sources of your interest, details of what you're looking for and the attributes that you'd like collected from them. That's all you need to be involved for. In the background, we custom crawl these data sources extracting record-level details as per your desired format. In addition, any template changes on the source site and other computational monitoring is taken care of through our platform.*

Key Words: Web browsing, web context, browser, online mode, offline mode, private vault.

1. INTRODUCTION

In the current era of online business, web based business have turned into a gigantic market for the general population to purchase merchandise online. Increasing utilization of savvy gadgets and different mediums has made ready for clients to purchase items nearly from anyplace. This has expanded association of online purchasers developing web based business. These substantial quantities of web based business sites place clients in turmoil to pursuit and purchase a solitary item from different web based business sites. The proposed arrangement helps online clients to get best arrangement for their item from various internet business sites on single web interface. This will thus spare clients time, cash and endeavors to locate a similar item costs on various internet business sites. Proposed framework utilizes web scratching method to concentrate information from web based business site pages furthermore web crawler to joins for items. This will likewise permit clients to examine costs and select items from same classification for contrasting its elements. When a data is searched, hundreds and thousands of results appear. The user's don't have persistence and stretch to go through each and every page listed. So, the search engines have a bigger job of sorting out the results, in the order of interestingness of the user within the first page of appearance and a quick summary of the information provided on a page. For instance, if the user is searching for cars, then the outcome returned are data about utilized cars available to be purchased instead of data about autos manufacture. Not all information represented is useful. For this we have to filter the query according to user interest.

2. SYSTEM ARCHITECTURE

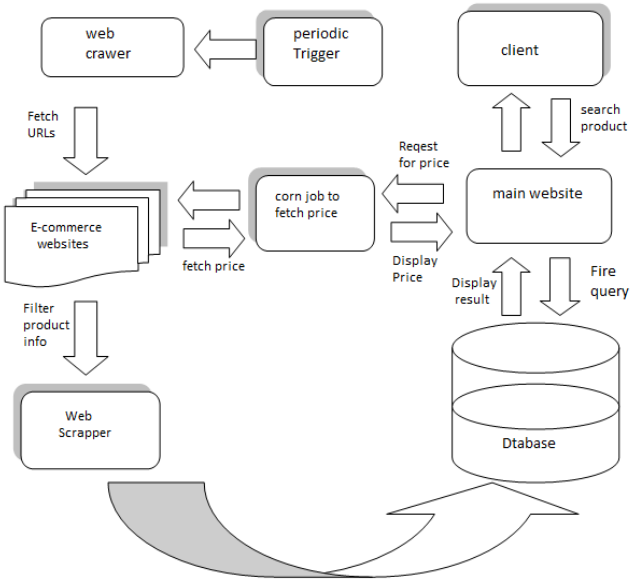

Figure1. System Architecture [Vengurlekar-et-al]

This system uses the following technologies:

1) Web crawler
2) Web scrapper

Also, we are going to use various technique in our application such as:

Classification technique

3. WEB CRAWLER

The system deals with price comparison engine .The first thing required are to gather large amount of data from different ecommerce websites. It is not possible to manually collect the data from websites. Hence the best way is to create a web crawler that will navigate to these e-commerce websites. The fetched URL's are send to scrapper for scrapping process.

The web crawler systems may get to be distinctly futile or junky if the data it draws are not pulling in users, particularly 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, vigor and the capacity to download huge number of pages. 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 as the results are in closed proximity of user's interest. The algorithm determines the relevancy based on the factors such as frequency and location of keywords.

3.1 Web Crawler Strategies :

1) Breadth First Search Algorithm

This algorithm aims in the uniform search across the neighbour nodes. It starts at the root node and searches the all the neighbour nodes at the same level. If the objective is reached, then it is reported as success and the search is terminated. If it is not, it proceeds down to the next level sweeping the search across the neighbour nodes at that level and so on until the objective is reached. When all the nodes are searched, but the objective is not met then it is reported as failure. Breadth first is well suited for situations where the objective is found on the shallower parts in a deeper tree. It will not perform so well when the branches are so many in a game tree especially like chess game and also when the entire path leads to the same objective with the same length of the path [2]. Andy yoo et al [3] proposed a distributed BFS for numerous branches using Poisson random graphs and achieved high scalability through a set of clever memory and communication optimizations.

2) Depth First Search Algorithm

This powerful technique of systematically traversing through the search by starting at the root node and traverse deeper through the child node. If there are more than one child, then priority is given to the left most child and traverse deep until no more child is available. It is backtracked to the next unvisited node and then continues in a similar manner [4]. This algorithm makes sure that all the edges are visited once breadth [5]. It is well suited for search problems, but when the branches are large then this algorithm takes might end up in an infinite loop [6].

3) Page Rank Algorithm

Page rank algorithm determines the importance of the web pages by counting citations or backlinks to a given page [7]. The page rank of a given page is calculated as

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

PR(A) ->Page Rank of a Website,

d ->damping factor

T1,....Tn ->links

Yongbin Qin and Daoyun Xu [8] proposed an algorithm, taking the human factor into consideration, to introduce page belief recommendation mechanism and brought forward a balanced rank algorithm based on PageRank and page belief recommendation which ultimately attaches importance into the subjective needs of the users; so that it can effectively avoid topic drift problems. Tian Chong [9] proposed a new type of algorithm of page ranking by combining classified tree with static algorithm of PageRank, which enables the classified tree to be constructed according to a large number of users' similar searching results, and can obviously reduce the problem of Theme-Drift, caused by using PageRank only, and problem of outdated web pages and increase the efficiency and effectiveness of search. J.Kleinberg [10] proposed a dynamic page ranking algorithm. Shaojie Qiao [11] proposed a new page rank algorithm based on similarity measure from the vector space model, called SimRank, to score web pages. They proposed a new similarity measure to compute the similarity of pages and apply it to partition a web database into several web social networks (WSNs).

4) Genetic Algorithm

Genetic algorithm is based on biological evolution whereby the fittest offspring is obtained by crossing over of the selection of some best individuals in the population by means of fitness function. In a search algorithm solutions to the problem exists but the technique is to find the best solution within specified time [12]. [13] Shows the genetic algorithm is best suited when the user has literally no or less time to spend in searching a huge database and also very efficient in multimedia results. While almost all conventional methods search from a single point, Genetic Algorithms always operates on a whole population. This contributes much to the robustness of genetic algorithms. It reduces the risk of becoming trapped in a local stationary point [14].

5) HITS Algorithm

This algorithm put forward by Kleinberg is previous to Page rank algorithm which uses scores to calculate the relevance [15].This method retrieves a set of results for a search and calculate the authority and hub score within that set of results. Because of these reasons this method is not often used [13]. Joel C. Miller et al [11] proposed a modification on

adjacency matrix input to HITS algorithm which gave intuitive results.

4. WEB SCRAPPER

Web Scrapping is utilized to concentrate HTML information from URL's and utilize it for individual reason. As this is value correlation site, information is scrapped from different online business sites. In this framework, Scrapping is done utilizing python libraries like solicitations and beautifulsoup4. Beautifulsoup4 is a python library which is utilized for parsing html pages. Utilizing these, item data from various internet business destinations is scrapped and put away in database. The span of the web is colossal; web indexes for all intents and purposes can't have the capacity to cover every one of the sites. Just 60 rates are the ordered web [11]. There is a high possibility of the important pages in the initial few downloads, as the web crawler dependably downloads website pages (in divisions). This calls for measures for organizing Web pages. The significance of a page is a component of its fundamental quality, its notoriety as far as connections or visits, and even of its URL. Distinctive specialists utilized diverse systems, for example, bread firth, profundity in the first place, page rank for selecting the sites to be downloaded. We need to begin from any URL (Seed), however envision that the beginning URL couldn't achieve all the website pages or even the pages referenced by seed URL doesn't reference it back, which in the end makes us to restart the creep. It is constantly better to have a decent seed URL – pages that have been submitted to them by dominant part clients around the globe. For instance Bing or Google. There is a cost connected with creeping, ordering and putting away the outcomes. At the point when the web gets greater and greater, the "better" pages are downloaded. So there should be a booking procedure to minimize creeping time and to decrease cost [4] and it varies starting with one web search tool then onto the next. As the web is immense and to download whatever number pages as could be expected under the circumstances, parallel crawlers are circulated so that numerous downloads can be done in parallel [5]

5. CLASSIFICATION TECHNIQUE

The information sets are gathered and prepared after which it is changed over into a reasonable arrangement. Post this progression, arrangement is done and results are shown in a justifiable format.[2] The second part in Fig 1 stores the E-trade information that will be characterized. The E-trade information has been gathered from prominent E-business locales. Well known datasets are utilized for the order of gadgets, for example, mobiles, pen drives and portable PCs. The third part display in Fig 1 manages the representation of the XML report. Pre-preparing is done utilizing the Document Object Model (DOM) Parser. The DOM parser is favoured over the Simple API for XML (SAX) parser as changes are not required in the XML information. Likewise, referencing back to the information is not required making DOM a practical option[8]. Figure 2 speaks to the yield got subsequent to applying the DOM parser to a XML record. The DOM parser gives the yield as a XML DOM tree. The qualities display in the dataset are vender rating, item organization, show no and cost. The DOM parser is built physically to parse these traits. The following part in Fig 1 manages changing over the XML information to an Attribute-Relation File Format (ARFF). ARFF records have two unmistakable segments. The primary area is the Header data, which is taken after the data. The ARFF is then nourished to the Waikato Environment for Knowledge Analysis (WEKA) instrument for characterization. The quantity of information focuses accommodated preparing are 70 while the test set has 30 information focuses.

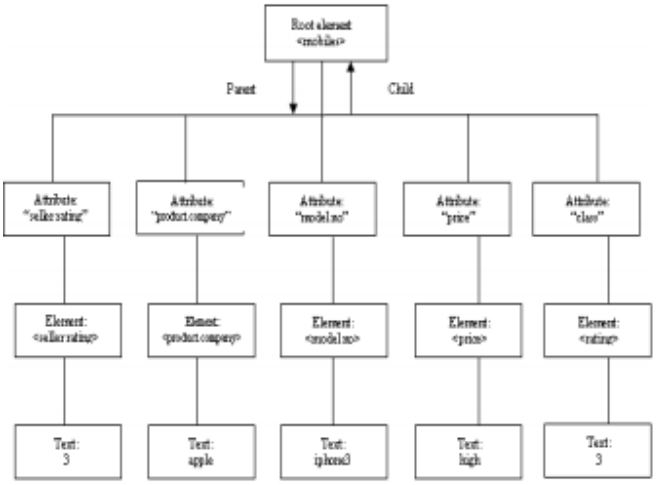

Figure 2. XML DOM Tree Example

The classification algorithm that is used is Naïve Bayes and Decision Tree. The methodology of classification for each algorithm is different and hence, gives us different results.

6. PROPOSED WORK

Working of the proposed system is as follows: The backend system consists of two important techniques web crawling and web scrapping and also classification technique, pattern matching algorithm and supervised algorithm is being. Web scrapping is a technique that is used to extract information in the human readable format and display it on destination terminal. But before scrapping the output, Web Crawlers are responsible to navigate to the destination once the crawler reaches the correct page and matches up with the products, scrapping process starts. Crawler periodically fetches information from e-commerce websites so as to check for updates. If updates are available crawlers carries those updates and makes necessary changes in the database, which will access in offline mode also and user can take a look for further update. Web scrapping essentially consists of two tasks: first is to load the desired web page and second is to parse HTML information of the page to locate intended information.

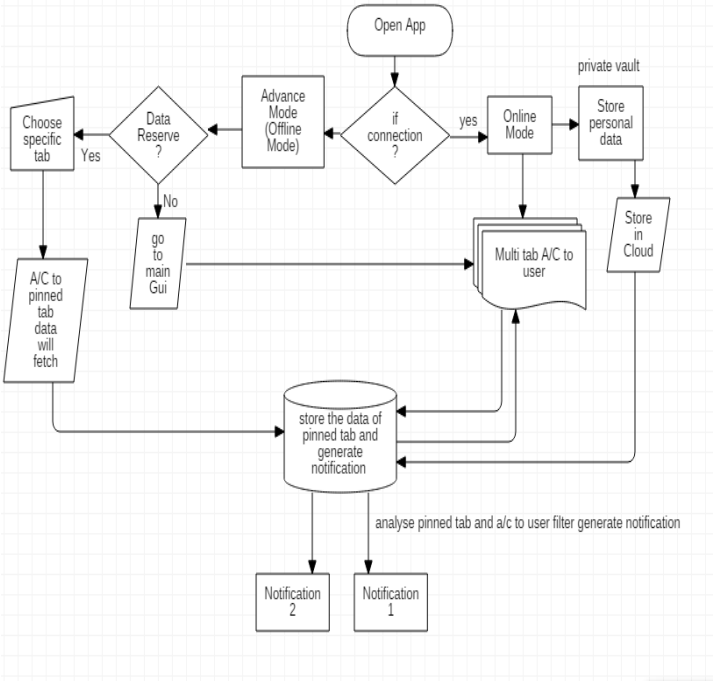

Figure 3. UML Diagram

7. CONCLUSIONS

The main objective of the review paper is for mining the relevant data among the all web-sites according to user interest for particular query and shows the best result using web crawling algorithms, classification algorithm and pattern matching. We moreover analyzed the distinctive request figuring and examine related to separate computations and their qualities and inadequacies related. We assume that most of the estimations concentrated on in this paper are effective for web looks for, yet the inclinations underpin more for Genetic Algorithm in view of its iterative decision from the people to make relevant results.

REFERENCES

[1] Pavalam S M, Jawahar M, Felix K Akorli, S V Kashmir Raja " Web Crawler in Mobile Systems" International Conference on Machine Learning (ICMLC 2011), Vol. , pp

[2] Wu Haitao, Tang Zhenmin "Automatic Classification Method for XML Documents" in International Journal of Digital Content Technology and its Applications (JDCTA) Volume5, Number12, December 2011

[3] Dr. Rajendra Nath ,Khyati Chopra," Web Crawlers: Taxonomy, Issues & Challenges" Proceedings of the International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3, Issue 4, April 2013, pp. 944-948

[4] Ricardo Baeza-Yates, Ricardo Baeza-Yates "Crawling a Country: Better Strategies than BreadthFirst for Web Page Ordering" , Proc. WWW 2005.

[5] Zailani Abdullah, Muhammad SuzuriHitam "Features

[6] Junghoo Cho and Hector Garcia-Molina "Effective Page Refresh Policies for Web Crawlers" ACM Transactions on Database Systems, 2003.

[7] Steven S. Skiena "The Algorithm design Manual" Second Edition, Springer Verlag London Limited, 2008, Pg 162.

[8] Ben Coppin "Artificial Intelligence illuminated" Jones and Barlett Publishers, 2004, Pg 77.

[9] Andy Yoo,Edmond Chow, Keith Henderson, William McLendon, Bruce Hendrickson, ÄUmit CatalyÄurek "A Scalable Distributed Parallel Breadth-First Search Algorithm on BlueGene/L" ACM 2005.

[10] Alexander Shen "Algorithms and Programming: Problems and solutions" Second edition Springer 2010, Pg 135

[11] Narasingh Deo "Graph theory with applications to engineering and computer science" PHI, 2004 Pg 301

[12] Sergey Brin and Lawrence Page "Anatomy of a Large scale Hypertextual Web Search Engine" Proc. WWW conference 2004

[13] Yongbin Qin and Daoyun Xu "A Balanced Rank Algorithm Based on PageRank and Page Belief recommendation"

[14] TIAN Chong "A Kind of Algorithm For Page Ranking Based on Classified Tree In Search Engine" Proc International Conference on Computer Application and System Modeling (ICCASM 2010)

[15] J.Kleinberg "Authoritative sources in a hyperlinked environment", Proc 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.