Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 5, May 2013, pg.317 – 322

RESEARCH ARTICLE

Analysis of Yslow Performance Test tool & Emergences on Web Page Data Extraction

N. Indira Priyadarsini[1], R. Mamatha[2]
[1]Vignana Bharathi Institute of Technology, India
[2]G. Narayanamma Engineering College for Women Shaikpet, Hyderabad, India

Abstract— Extracting the data from web is a vital role analysis for several web data applications, slow performance of webpage has many issues. Previous work shows on data extraction techniques in multiple page level. In our work we introduced the concept of Yslow testing process for webpage reduces the issues such as HTTP request, response & Domain Network System, JavaScript and also improves page level data extraction performance in positive rates. We also propose best testing tools study on webpage increases extracting speed and resolve the JavaScript, cascading style sheet.

Key Terms: - Webpage application; Yslow; Http request & response; Performance testing tools

I. INTRODUCTION

Extracting the knowledge or relevant data in web is a technique of various web resources to collect required information enables to an individual or organization to promote business understanding marketing dynamics. Complex unstructured data are appearing on the web, automatic systems are generated for web data extraction techniques such as price comparison and recommendation systems. Web page contains valuable information than the surface web to consolidate information requires substantial efforts the pages are generated for visualization not for data exchange. Generating an extraction program for a given search form is equivalent to wrapping a data source such that all extractor or wrapper programs return data of the same format. Merging the data means for example will keep data updates in variety of devices no longer access to single computer but several in the form of home and work station mobile phones, having information in several places leads how to keep it in synchronized for this purpose use the margin in web data extraction. To extract data from web template as a common model for all webpages, template contains alignment text HTML tags to highlight the text requires hypertext are the part of template produces many equivalent classes making the reconstruction of the schema. This work focus on the relevant multiple page level extraction using FivaTech to automatically detect the schema of the website that carry the knowledge information called pattern mining as well as the tree template. Web data extraction wrapper is used in languages for wrapper development HTML aware tools modeling based tools and ontology based tools. Some information is fit in two or more identified groups. The development of language specially designed to assist construction of wrappers such languages were proposed as general purpose languages such as Perl and Java. Inherent structural features of HTML documents for data extraction performing the extraction process tools turn the document into a parsing tree a representation that reflects its HTML tag hierarchy representative tools such as an approach are W4F, XWRAP and Roadrunner. Natural language processing techniques used by several tools to learn extraction rules for extracting relevant data existing in language documents tools apply techniques such as filtering part of speech. NLP tools are usually more suitable for webpages consisting of free text in telegraphic style as job listings are RAPIER SRV and WHISK. Wrapper induction tools generate delimiter extraction rules derived from a given set of training examples tools more suitable for HTML documents than the previous ones. Ontology tools rely on the structure

of presentation features of the data within a document to generate rules or patterns to perform extraction. Extraction can be relying directly on the data to locate constants present in the page and to construct objects.

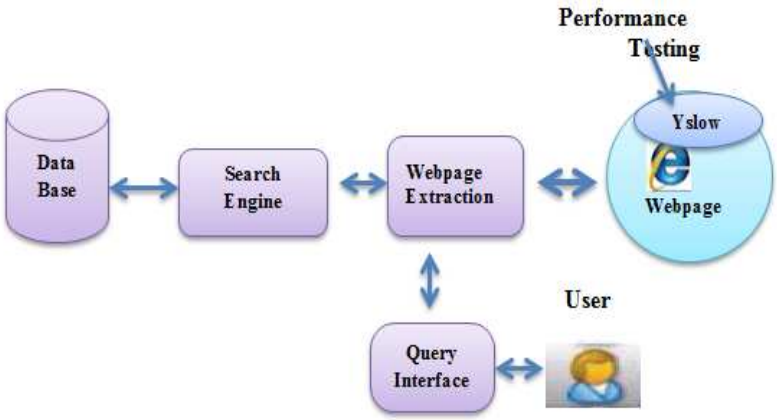
II. RELATED WORK

Mozenda: Process of pulling information from designated internet webpages where we want to store and find it help to retrieve relevant information. Over time we can extract web data from multiple locations and save it to spread sheets of XML documents Mozenda extracts data from multiple sources and combines it in a manner that enables users to have useful update information wherever need it. Mozenda is software as a service company that enables users of all types to easily and affordably extract web data. Users can set up agents that routinely extract data store data and publish data to multiple destinations. Once information is Mozenda systems user can format repurpose and mashup the data to be used in other online application as intelligence. It is secure hosted in class a data warehouse but can be accessed over the web securely the web console.

A webcrawler is software written in programming language such as Java made up of several lines of code that give instructions on what to search and designed to search for certain types of information on webpages such as keywords meta tags hyperlinks and descriptions. Webcrawler software to search the web for audio video files any type of data or text files and used to browse the internet in a systematic mechanical manner. Many search engines such as Google use this to provide current data. Webcrawler is used to create a copy of sites visited on the internet and processed by the search engine then will index the pages that have down loaded for faster searches. Webcrawler is also used to perform automated maintenance tasks on websites. The process of performing a website data scrape involves using computer software to extract specific information from a number of websites, it written simulates the process by which a human would search for information. Website scraper is the process of performing using software to extract specific information from a number of websites. This data is collected for a number of reasons at first to facilitate the process by which companies would monitor online competitor pricing. The technology has been growing, and has been made available to a wider variety of users in an affordable fashion, such as the Mozenda website data scrape software. Such services now allow users and companies to perform data scraping without extensive knowledge of computer programming. Website scrapers are useful in many professions. For example, realtors may need to find a house without a mouse. A detective with mace for a case in haste puts the scraper to a caper without need for pen or paper. For those in health, this information equals wealth.

III. PROBLEM DEFINITION

Many performance issues like HTTP Request, JavaScript, and Domain Network System of webpage describe the slow access when we retrieve the data from website. Webpage is created by embedding the data object into a predefined format, to resolve all these page became complex. To increase the performance of webpage our work analyzes performance testing using Yslow and HTTP watch. Slow Cascading Style sheet performance causing slow running JavaScript is slow in internet explorer; the developer may use lookup methods via CSS provided by JavaScript frameworks such as jQuery or prototype. It needs to iterate the whole DOM tree to perform the lookup purely through JavaScript.



IV. PROPOSED FRAMEWORK

According to problem analysis this framework for extracting information from web sources and implementing the performance testing to increase webpage data extraction. Query interface is the key word for accessing information, user writes the name of topic, information or any product name, query interface sends to a search engine which searches the web data sources, the results of the query will be saved as page by query interface as HTML files.

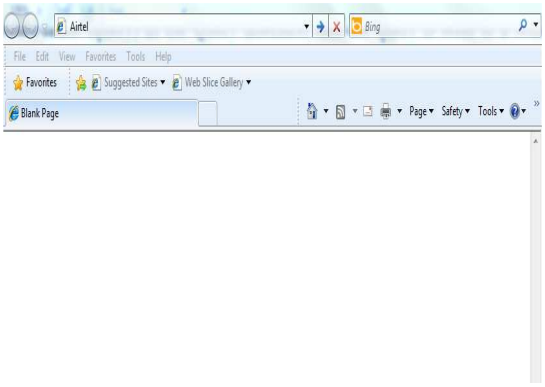


Figure 2

The result of a query and the web documents which are stored as webpage for particular keyword

When retrieving the information on query based keyword sometimes process is very slow, to increase the performance of webpage our analysis used best tools Yslow.

V. YSLOW

Yslow is the yahoo's development and analyzes web page by examining all the components on the page dynamically created by using JavaScript, Components, Statistics and Tools grading is done on 22 performance metrics divided into six categories content cookies CSS images JavaScript and server. The components provides all components that YSlow detects web page uses and lists very useful and pertinent information about them under five categories doc HTML, XML or XHTML JavaScript CSS statistics compares the page loading when components were not cached verses when they components were actually cached run the Yslow test, tools provide features that can make some optimizations much easier to implement suggestion from previous post then click all CSS link to combine used by web page. Performance speeding up by the following

5.1 HTTP Requests: Time spent on 80% end-user response most of this tied up in downloading all the components in the page. Images style sheets, scripts flash etc. reducing the number of components in turn reduces the number of HTTP requests required to render the page, it is the key to fast pages. Combined files are a way to reduce the number HTTP requests by combining all scripts into single script and similarly combing all

CSS into a single style sheet. CSS sprites are the preferred method for reducing the number of image requests combine the images into single image and use CSS background image position properties to display the desired image segment.

5.2. Use of JavaScript and CSS External: Using external files generally produces faster pages because the JavaScript and CSS files are cached by the browser, JavaScript and CSS are Inlined in HTML documents every time the HTML document is required. This reduces the number of HTTP request that are needed but increases the size of the HTML document other one if the JavaScript and CSS are in external files cached by the browser the size of the HTML document is reduced without increasing the number of HTTP requests. Many websites fall in the middle of these metrics; the best solution generally is to deploy the JavaScript and CSS files exception where inlining is preferable with homepages such as Yahoo front page and My yahoo. For front pages that are typically many pages views are techniques that leverage the reduction of HTTP requests that inlining provides as well as caching benefits achieved through using external files.

5.3. Reduce DNS Lookups: The domain name system maps hostnames to IP addresses just as phonebook map people names to their phone numbers when we type link in the browser, a DNS resolver contacted by the browser returns that servers IP address DNS has a cost typically takes 20-120 milliseconds for DNS to lookup the IP address for a given hostname. DNS lookups are cached for better performance maintained by the users ISP or local area network but there is also caching that occurs on a special caching server maintained by the users ISP or local area network but there is also caching that occurs on the individual user computer. The DNS information remains in the operating systems DNS cache on Microsoft windows.

Internet explorer caches DNS lookup for 30 minutes by default as specified by the DNS cache time out registry setting. Firefox cache DNS lookup for 1 minute controlled by the network, DNS Cache Expiration Configuration Setting when the client DNS cache is empty number of DNS lookups is equal to the number of unique hostnames in the webpage used in the pages URL images, script files Style sheets Flash objects. Reducing the number of unique hostnames has the potential to reduce the amount of parallel downloading that takes place in the page. Avoiding DNS lookups cuts response times but reducing parallel downloads may increase response times.

5.4 Remove Duplicate Scripts: A review done by US websites that two of them contain a duplicated script. Two main factors increase the odds of a script duplicated in a single web page team size and number of scripts when it does happen duplicate scripts hurt performance by creating unnecessary HTTP requests and wasted JavaScript execution. In addition to generate wasteful HTTP requests time is wasted evaluating the script multiple times this redundant JavaScript happens in both Firefox and internet explorer.

5.5 Use Cookie free Domains for Components: When the browser makes a request for a static image and sends cookies together with the request the server doesn't have for those cookies so they only create network traffic for no good reason. We make sure static components are requested with cookies free requests. If domain is www.forexample.org then all requests to static example will include those cookies free Yahoo uses other benefits of hosting static components on cookies free domain is that some proxies might refuse to cache the components that are requested with cookies.

VI. STUDY ON PERFORMANCE TOOLS ON WEBPAGE

Several website performance tools loading speed into consideration in determining its ranking, this will not land a big impact in other words website is slow we are subjected to lose a lot of visitors are traffic, traffic affects reputation. One of popular Firefox performance add-ons Yslow and Google page speed gives suggestions what we need to improve in our webpage are recommended for Firefox can also be applicable to other browser. Site-Perf emulates natural browser downloading page with all the images CSS JS and other files like regular visitor on the report can see website page loading what files start loading at first and how fast it is, very useful performance report to find elements that need to improve website loading time. WebToolHub gives the option to understand how page is loading with different visitors internet connection speeds get the information of page size loading time with different internet connection speeds all about CSS. Google page speed is an open source Firefox add-on for webmasters and web developers can use page speed to evaluate the performance of their webpage and get suggestions on how to improve. Comparing to all webpage test tools Yslow user needs to install Firebug add on for Firefox first and then install the Yslow add on. In order to run Yslow and view the results Firebug must be enabled and the Firebug window must be open user can right click on the Yslow icon and check the auto run feature. Features of Yslow user can choose any of the predefined rule set like Yslow small site or blog to rate the website on those set of parameters, which provides an option to the user to filter the parameters based on content cookies CSS Image JavaScript Server. It groups the components under different

subgroups like doc, js, css, iframe, cssimage and image, shows the total number of HTTP request and a total weight in terms of empty cache and primed cache as a weight graphs.

It gets all the components of the page by crawling the DOM, fetches information such as size whether it was gzipped expires header etc. If the component information is not available from Net panel for example component read from cache or it had 304 response YSlow makes an XMLHttp Request to fetch the component and track its header and other necessary information, Yslow generates a grade for each rule taking into account all this data of the page and ultimately produces the overall grade for the webpage.

The user can significantly improve the performance of the website
The user can design faster webpages
The user can reduce the end-user response times
The user can get the most potential for improve by focusing on the front-end.

VII. CONCLUSION

In this paper, presents multiple page-level web data extraction, A framework Yslow tool control delays in web page extraction, along with Yslow some more tools specification mechanism shows resolves the issues in page level and increases best performance in web application. Work analysis, how we enters the request in webpage and retrieves the data from web server has been discussed as well, followed by the study on efficient webpage test tools evaluation of our method. Future work extends to investigate more analysis services for web page mining with novel research.

REFERENCES

[1] B. Chandrasokaran, John R. Josophson, and V. Richard Bonjamins (1999). What are Ontologies, and Why Do We Need Them?, Journal of IEEE Intelligent Systems and Their Applications, Vol. 14, Issue. 1, pp. 20-26.

[2] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, and Khaled F. Shaalan (2006). A Survey of Web Information Extraction Systems. Journal of IEEE Transaction on Knowledge and Data Engineering, Vol. 18, Issue 10, (Oct. 2006) pp. 1411-1428, ISSN: 1041-4347.

[3] David Buttler, Ling Liu, and Calton Pu. 2001. A Fully Automated Object Extraction System for the World Wide Web, Proceedings of the 21st International Conference on Distributed Computing Systems, Georgia Institute of Technology, ICDCS, pp. 361-370, ISBN: 0-7695-1077-9, 2001, USA..

[4] Domenico Beneventano and Stefania Magnani (2004). A Framework for the Classification and the Reclassification of Electronic Catalogs, Proceedings of the 2004 ACM Symposium on Applied Computing, pp. 784-788, ISBN: 1-58113-812-1, Nicosia, 2004, Cyprus.

[5] Guntis Arnicans and Girts Karnitis (2006). Intelligent Integration of Information from Semi-Structured Web Data Sources on the Base of Ontology and Meta-Models, Proceedings of the 7th International Baltic Conference, pp. 177-186, ISBN: 1-4244-0345-6, Vilnius, July 2006, Latvia University, Riga.

[6] Horacio Saggion, Adam Funk, Diana Maynard, and Kalina Bontcheva (2008). Ontology based Information Extraction for Business Intelligence, In: Lecture Notes in Computer Science, pp. 843-856, Springer Berlin, Heidelberg, ISSN: 0302-9743 (Print) 1611-3349 (Online).

[7] Jeong-Woo Son, Jae-An Lee, Seong-Bae Park, Hyun-Je Song, Sang-Jo Lee, and Se-Young Park. 2008. Discriminating Meaningful Web Tables from Decorative Tables using Composite Kernel, Proceedings of ACM International Conference on Web Intelligence and Intelligent Agent Technology, Vol. 1, pp. 368-371, ISBN: 978-0-7695-3496-1.

[8] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, Reading, Mass., 1999.

[9] S. Chakrabarti, Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data, Morgan Kaufmann, San Francisco, 2002.

Authors Bibliography

N. Indira Priya Darsini
M.Tech Information Technology from Gurunank Engineering College B.Tech Computer Science Engineering from Bojji Reddy College of Engineering. She is having seven years of academic experience currently working as Asst Prof at Vignan Bharathi Institute of Technology, she has guided many UG & PG students. Her research areas include Multimedia Web Technologies, Testing, Databases and published two articles in International Journals

R. Mamatha
M.Tech Software Engineering Nishitha Engineering College M.B.A from Osmania University B.Tech from Greenfort Engineering College. She is having seven years of Academic experience currently working as Asst Prof at Narayanamma Engineering College, she has guided many UG & PG students. Her research areas include Multimedia Web Testing, Web Mining, Data mining

