International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.2, 2015

RESEARCH ARTICLE

ISSN: 2321-7758

EFFECTIVE SOLUTION ON DATA FROM MULTIPLE WEBSITE USING EXTRACTION AND DATA ANALYSIS TECHNIQUE

D.DHAYALAN[1], BHUVANA.V[2]

[1] Assistant Professor, [2] PG Scholar, Department of MCA

[1,2]Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Avadi

Article Received: 11/04/2015 | Article Revised on:15/04/2015 | Article Accepted on:18/04/2015

D.DHAYALAN

ABSTRACT

The data from the different website can be extracted using the web data extraction method. The website used for this extraction is in the similar form. This technique fetching the data's from the different website and removes the unwanted stuffs from the data's and provide the user needed data's in a single window with best result. In existing the data will be extracted from a single website and provide the result of a single window, when the comparison of multiple website made, it uses multiple window, so the comparison of different websites will be more difficult. So the proposed system used to extract data's from multiple websites and provide result in a window the comparison can be made easily. It provides an exact best result among the different websites and also it is also more efficiency to use. Here we are using the data extraction, Stemming method and exploratory data Analysis method to implement the process. The Data extraction method is used to extract the data from the multiple websites that is used for same purpose and the Stemming method is used to remove the unwanted content in the multiple extracted data finally the Exploratory data analysis method is used to analysis the data that is gotten from the stemming process and produce the best result among them. The main purpose of using this technique is for comparing the website that is used for an on-line shopping, here it provide the best website for a purchasing.

Keywords: Data extraction, Stemming process, Analysis Method, wrapper generation, Web extraction, Web mining.

1 INTRODUCTION

There are many website can occur for single purpose, at that time the user cannot know which is the best website to work on it, so the user waste their time to searching for an different websites and finally they take decision physically, sometime the decision cannot be right, so the technique used to make the comparison between multiple website an provide best result. Some technique extract the data from the single website then remove the unwanted data and provide the data alone to the user, here the comparison can be made by multiple window as

Hassan A. Sleiman says[1]. Some technique use to compare the HTML pages alone and produce the automatic wrapper as Valter Crescenzi says[9]. The path clustering is used to extract the multiple website data and provide it most effective, it provide a multiple data but not Analysis as Gengxin Miao says[10]. Some methods uses the automatic web data extraction method to extract the data from the websites but it provide all the data's in the reportas Devika.K says[12]. Many techniques uses the tree shape techniques that are used to manipulate the extracted data's and filter the wanted data alone and produce the resultas Yanhongzhai says[4]. The region extractor can be used in some technique to survey the process as Hassan A.Sleiman says[6]. In Existing many technique uses many methods and technique to extract the data but uses for different purpose. It does not provide any tools for extracting data's from the multiple websites and provide it in a single window and Analysis the data and finally it does not produce the report. So the proposed system uses the extraction, stemming, analysis methods to provide the comparison result among the multiple websites.

2 General Views

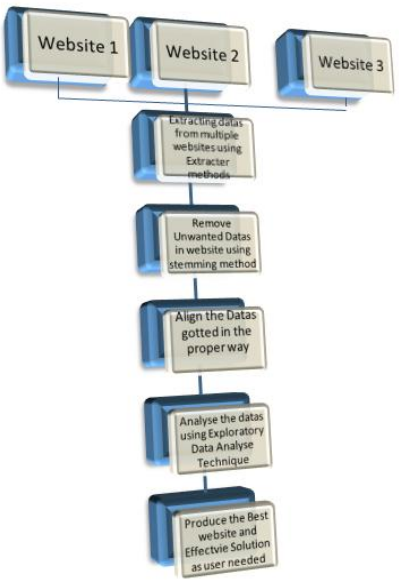

Fig 1 General View web extractor and Analysis method.

The Fig 1 shows the general view of the effective solution on data from multiple website using Extraction and Analysis technique, as can be seen from the figure, the web extractor extract the data from the multiple websites(website 1, website 2,... website n). The extractor will extract the source code of the every website that has to be compared. Then the unwanted data's in the each website can be removed using the stemming method. The stemming method will filter the extracted source code and provide the needed data for comparison. It removes the unwanted stuff. Then the data had gotten after stuffing process will aligned and provided in a single window. After this, the Exploratory Data Analysis technique will be used for Analysis the data. The data extracted from the different website can be analysis, compare and provide the best data's or a best website among them. This technique can be help to Analysis the on-line shopping website to shop in the best website. Finally, it provides the result and report as a best website among the extracted websites.

3 Earlier functions of Extracting and analyzing data on the website

The Existing System extracts the data form single website and provides it in a single page. The remove of unwanted data is very difficult in the existing system. Some time the extracted data will be lost due to some confusion. The comparison of multiple website will be cannot made here. It is possible to compare the websites with help of two or more window at a time. It cannot provide a proper conclusion among the product provided in the website. It takes more time to compare the multiple websites. The methods used in the existing will take more time to evolve so we propose the technique Data extraction, Stemming method and exploratory data Analysis methods. In proposed we can compare the multiple website easily.

4 Present Scenarios of Data Extraction and Analyzing the Multiple Websites

In this study, it is comparing the multiple website and Analysis it. The existing system will extract the data from the single website and provide the data in a single window. The comparison of different website will be difficult, so the proposed system uses, the data from the multiple website will be extracted and provide the user needed data alone in a single perspective. Then it makes the Analysis of data from the multiple website and provides the best data occurs in a best website, the technique used in the proposed is the Data extraction,

Stemming methods and then the exploratory data Analysis methods. This technique use to compare the multiple websites. The data extraction is the method used to extract the data from the different websites. Then the Stemming process is used to remove the unwanted data occurs in the websites and provide the user needed data's alone. The exploratory data Analysis technique is used to Analysis the data that got after stemming process the best data among them and also provide the best website. Finally the report or result will be generated and provide to the user as a best solution.

5 Methodologies

5.1 Data Extraction Method

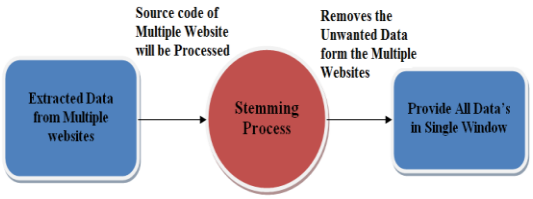

Fig 3 Process of Stemming Method

The Fig 3 represents the process of Stemming method. The Data extractor will extract the source code and data of the multiple websites and the stemming process will removes the unwanted data from each and every websites and provide user needed data in a single window to compare the different websites in a single window.

5.3 Exploratory data Analysis

This is the technique used for Analysis the several data. The data extractor extract the data's as a source code from the websites and the stemming process will be removes the stuff occurs in every websites. Then the Exploratory data Analysis technique will used to Analysis the data that is extracted from the different websites and provide the final result that, which website will be the best website among the multiple websites. The Exploratory data Analysis technique will Analysis the each and every data or field occurs in the website and provides the result. Finally it produces the report that, which is the best website. This technique will provide the user needed suggestion among the multiple websites. The statically model of Exploratory data Analysis is shown in the below equation.

$$p(y_i|\mu_1,\mu_2,\sigma_1,\sigma_2) = .5\frac{1}{\sigma_1}\phi\left(\frac{y_i-\mu_1}{\sigma_1}\right) + .5\frac{1}{\sigma_2}\phi\left(\frac{y_i-\mu_2}{\sigma_2}\right),$$

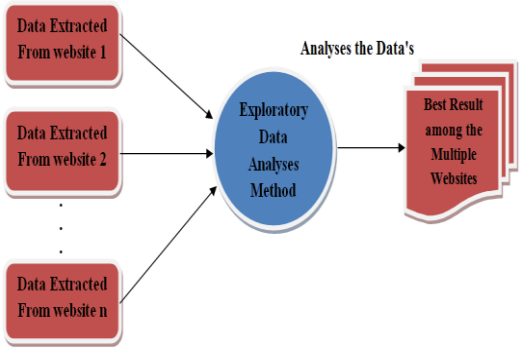

Fig 4 Exploratory data Analysis Technique

The Fig 4 represent the Exploratory data Analysis Technique, Here the extracted data from the website 1, website 2,…, website n will be analyzed using the Exploratory data Analysis Method. And finally report the result to the user that the best Website among the Multiple website that is extracted. It provide the best Suggestion to the user about the best website among the multiple websites.

6 Simulation Results

The simulation results shows about the extraction of data from the websites and the Analysis can be made and provide the user needed data in a window. Here we simulate the existing and the proposed system for extracting the data's from the website for the user convenient.

The Fig 5 shows the simulation result of the existing system and the proposed system used for extracting and analyzing. Here, the existing system does not extract the data's from the multiple website so, the multiple extraction cannot be made but in proposed system, the technique can extract the data from the multiple websites and provide that data for processing so here the multiple extraction can be made easily. Comparing to the existing system, the proposed system remove the unwanted data's more which occur in the each and every website that is extracted. Then the analyzing data's can be made, the existing system contain the single website data alone so it cannot Analysis the data more, but in the proposed system it will Analysis each and every data's or a field that occurs in different website and provide the result.

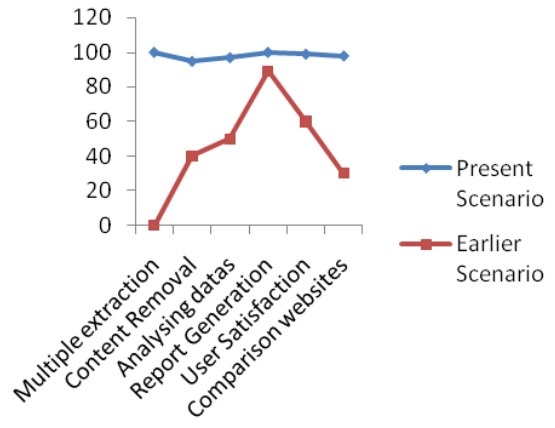

Fig 5 Simulation result of existing and proposed system of web extractor and Analysis method

The Report Generation can be made at the final stage, both existing and proposed system will produce the report but the existing provide the data occurs in the single website and the proposed system provides the data occurs in the multiple website can be presented in a single window, here the comparison can be easily made. The existing system are does not provide the effective solution to the user but in proposed the effective solution can be provided to the user so it is more user satisfaction. Then the comparison of multiple website be difficult in the existing system, it uses the multiple window for comparing the different website but here the proposed system uses the different technique to extract the data's from the different website and provided in a single window, so the effective comparison can be made. From the result of the simulation result the proposed system will be more effective and efficiency to the user.

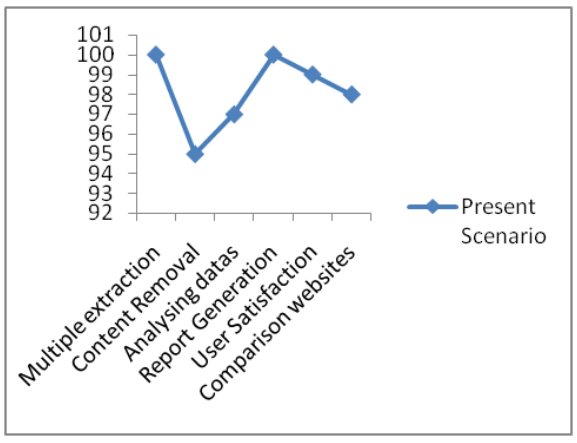

Fig 6 Present Scenario of Simulation Result

The Fig 6 Shows that the Percentage usage of Multiple extraction, Content Removal, Analyzing data's, Report Generation, User Satisfaction an Comparison websites of the present Scenario. All the usage of the Present Scenario will be more than the 90 percentage; it shows the present scenario is the best among the others.

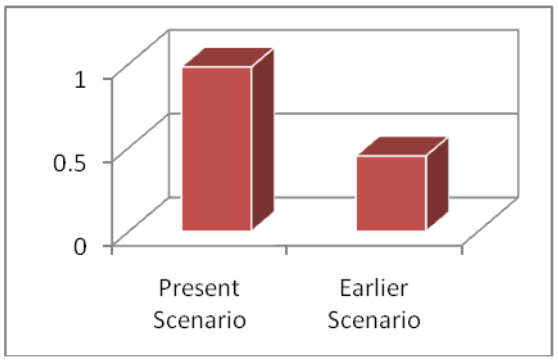

Fig 7 Overall Percentage of Present Scenario an Earlier Scenario

The Fig 7 represents overall percentage of the present scenario and the earlier scenario of the data extraction and analyze methods on the websites. From the Bar Chart it shows that the Present Scenario will hold 98% and the Earlier Scenario Will hold 45%, from this the present scenario will be the best, and it provide the best result to the user among the multiple websites.

Conclusion

The work presented in this paper is extracting the data's from the multiple website and removes the unwanted data's from the websites and uses the Analysis method to Analysis the data and provide the result to the uses. The result will be more effective than the other. Here the multiple website data can be provided in a single window. This makes the user to compare the multiple website at a time. It reduces the time and provides the user need. The comparison of different website can be easily made in this process. Here we are using the extraction, stemming, exploratory data analysis methods to complete the process properly as per the user satisfaction.

For future scope, this technique can be implemented everywhere not only in the purchasing websites. It can be implemented for every fields to Analysis the data.

REFERENCE

[1] Hassan A. Sleiman and Rafael Corchuelo, "Trinity: on using trinary trees for unsupervised web data extraction", ieee

transactions on knowledge and data engineering, vol. 26, no. 6, june 2014.

[2] C.-H. Chang and S.-C.Kuo, "Olera: semisupervised web-data extraction with visual support," *ieee intell. syst.*, vol. 19, no. 6, pp. 56–64, nov./dec. 2004.

[3] Chia-Hui Chang, Member, ieee computer society, mohammed kayed, mohebramzygirgis, member, ieee computer society, and khaled f. Shaalan, "A survey of web information extraction systems", ieee transactions on knowledge and data engineering, vol. 18, no. 10, october 2006.

[4] YanhongZhai and Bing Liu, "Structured data extraction from the web based on partial tree alignment", ieee transactions on knowledge and data engineering, vol. 18, no. 12, december 2006.

[5] Minky Jindal and NishaKharb, "K-means clustering technique on search engine dataset using data mining tool", International journal of information and computation technology.

[6] Hassan A. Sleiman and Rafael Corchuelo. " A survey on region extractors from web documents", ieee transactions on knowledge and data engineering, vol. 25, no. 9, september 2013.

[7] D.PramodKrishna, T.SwarnaLatha and T.Rajasekhar Reddy, "Extracting web data based on partial tree alignment using fivatech", International journal of advanced research in computer science and software engineering volume 2, issue 3, march 2012 issn: 2277 128x.

[8] NeerajRaheja and V.K.Katiyar, "Efficient web data extraction using clusteringapproach in web usage mining", IJCSI international journal of computer science issues, vol. 11, issue 1, no 2, january 2014.

[9] ValterCrescenzi, Giansalvatore Mecca, Paolo Merialdo, "Roadrunner: towards automatic data extraction from large web sites".

[10] Gengxin Miao, Junichi Tatemura, Wang-Pin Hsiung, ArsanySawires, Louise E. Moser," Extracting data records from the web using tag path clustering".

[11] Kyle Williams, Lichi Li, MadianKhabsa, Jian Wu, Patrick C. Shihand C. Lee Giles, "A web service for scholarly big data information extraction", 2014 ieee international conference on web services.

[12] Devika K, SubuSurendran, "An overview of web data extraction techniques", International journal of scientific engineering and technology (issn : 2277-1581) volume 2 issue 4, pp : 278-287 1 april 2013.

[13] C.-N. Hsu and M.-T. Dung, "Generating finite-state transducers for semi-structured data extraction from the web," Inform. Syst., vol. 23, no. 8, pp. 521–538, Dec. 1998.

[14] P. Gulhane, R. Rastogi, S.H.Sengamedu, and A. Tengli, "Exploiting content redundancy for web information extraction," in Proc. 19th Int. Conf. WWW, Raleigh, NC, USA, 2010, pp. 1105–1106.

[15] N. Kushmerick, D. S. Weld, and R.B.Doorenbos, "Wrapper induction for information extraction," in Proc. IJCAI, 1997, pp. 729–737.

[16] I. Muslea, S. Minton, and C. A. Knoblock, "Hierarchical wrapper induction for semistructured information sources," Auton. Agents Multi-Agent Syst., vol. 4, no. 1–2, pp. 93–114, Mar./Jun. 2001.