Contemporary Engineering Sciences, Vol. 8, 2015, no. 11, 491 - 497
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2015.53128

Automated Trinity Based Web Data Extraction for Simultaneous Comparison

A. Viji Amutha Mary[1], S. Justin Samuel[2] and D. Johnny Rajam[3]

[1] Dept. of CSE, Faculty of Computing
Sathyabama University, Chennai, India

[2]Dept. of IT, Faculty of Computing
Sathyabama University, Chennai, India

[3]Department of Computer Science
Sathyabama University, Chennai, India

Abstract

Most of the users search for an effective system which can provide an optimized comparative solution for online shopping. Trinity is an innovative framework to extract data from the internet based web applications. It subdivides the web patterns into smaller pieces of patterns which includes prefix, suffix and separators. And the data will be cleaned and formatted for manipulation which gives an efficient cost comparative system.

Keywords: Web data extraction

1. Introduction

Online shopping is the new trend world wide. Shopping online is, trading products using networks such as the Internet. Though online shopping has many Cons it also has lots of Pros. Unlike buying a product in a local store, buying a product online won't consume lots of time. People prefer online shopping because it is available all the time (24/7), more convenient, can avoid crowding, possibility

to buy used products in lower price, lots of varieties, can view the reviews of customers, can compare prices of the product, etc.

There are masses of websites available to buy a product Online. Generally a shopping website is rated based on the offers and discounts, reviews, reasonable cost, delivery time, mode of payment, shipping charges, etc. Due to the ample shopping websites the challenging task for the users is to determine which one to prefer. And also to make sure that they have visited the best website for buying a product. The web sites should be compared and the comparison involves the cost of the product. This comparison will allow the user to select the best website.

2. Related Work

Hassan A. Sleiman and Rafael Corchuelo pictures that an unsupervised technique to extract data from web pages. The shared pattern doesn't give relevant data, therefore, split it into prefix, separator and suffix and generates a regular expression which represents the template [1]. Arasu and H. Garcia-Molina explain that the pages created from an unknown template and values can have multiple solutions. Words that are part of template have a high correlation of occurrence in the input page with others words [2]. The goal is to choose the correct template and identifying the template and the attributes correctly. J. L. Arjona, R. Corchuelo, D. Ruiz, and M. Toro in "From wrapping to knowledge" reveals that It extracts the attributes from the web page. The knowledge model represents it by class & roles (semantics). The proposal allows transferring wrapper's o/p into knowledge entity [3].

F. Ashraf, T. Özyer, and R. Alhajj in "Employing clustering techniques for automatic information extraction from HTML documents" uses clustering for data Extraction [4]. The proposed approach is very simple and instead of classifying tokens individually, we can extract all tokens in one go. C.-H. Chang and S.-C. Kuo reveals that the users can select data of interest before training process and name the extracted data after the system generalizes the extraction rules. The only harm can be the extraction failure [5]. Yanhong Zhai and Bing Liu in "Web Data Extraction Based on Partial Tree Alignment" define the identification and Alignment of data to extract the exact data [6]. Hazem Elmeleegy, Jayant Madhavan and Alon Halevy tells that the ListExtract is one of the unsupervised methods which extract quality tables from web [7]. C.-H.Chang, M. Kayed, M. R. Girgis, and K. F. Salon explains that the algorithm used

experience a replicated slip-up data with extensive variations in the parameters of the supremacy system network, counting noisy environment,providing a reliable measure of 99% with a quicker response time. It finds to be fast and very accurate [8]. Valter Crescenzi and Giansalvatore Mecca put forward a novel multitask spectrum-sensing process on the basis of spatiotemporal data mining method. Within every cluster, we suppose with the intention of spectrum sensing is executed in a synchronized technique [10]. The Dirichlet process preceding is engaged to build a routine confederacy of the spectrum sensing consequences amongst diverse clusters with a widespread sparseness hyper parameter communal within every group.

A. Carlson and C. Schafer used a two-step data mining-based method for plastic bearing fault diagnostics by the means of vibration sensors. The frequency domain CIs are used through a statistical classification model to recognize bearing outer race faults [9]. The time domain CIs extracted using EMD are urbanized to build a k-nearest neighbor algorithm- based fault classifier to support other types of bearing faults.

3. Proposed System

The Proposed system focuses on designing a multi perspective, crawling mechanism for fetching the information from multiple websites. An automated stemming process is used to remove the unwanted data after fetching the website structure. And an automatic manipulation takes place and the data will be formatted based on user requirement. The comparative analysis gives the best solution for the buyers. It also uses multiple features for comparison.

3.1 ALGORITHMS USED

- Dominant Superstring Algorithm
- Ant colony optimization Algorithm

3.2 ADVANTAGES

- Effective cost comparison.

3.3 TECHNIQUES USED IN PROPOSED SYSTEM

- Crawling Website
- Extraction of prefix and suffix
- Irrelevant content removal
- Consolidating website content
- Recommendation analysis
- Performance Evaluation module

3.3.1 Crawling Website

Uses web crawling or spidering technique in order to crawl the website. The user can crawl 2 to 3 pages simultaneously. The Multi-threaded Web crawler will be very efficient. Fig 1 represents the framework of the proposed system. Fig 2 shows the screen shot of parallel crawling of the website.

3.3.2 Extraction of prefix and suffix

This module enables the access point to extract all the Links and Sub links. The contents of the URL will be extracted as prefix, suffix and separators**.** The Prefix contains the head section and the Suffix will have the Script in the Webpage. And the Separators will have the Body content.

3.3.3 Irrelevant content removal

This module is used to neglect the unwanted URL's and the contents from the website. Irrelevant content removal module will remove the irrelevant contents. This module consolidates the analyzed websites and generates the report. This report is nothing but the extracted information from the relevant and irrelevant contents are available in the website

3.3.5 Recommendation analysis

Recommendation zone analysis module tends to recommend a systematic assessment of specific website zone content and impacts the extraction of search.

3.3.6 Performance Evaluation module

In the performance evaluation module consolidated, analyzed and identified report will be submitted which will relay on the performance of relevant search of the document. In Fig 3, the product statistics is shown clearly through a cost comparison chart.

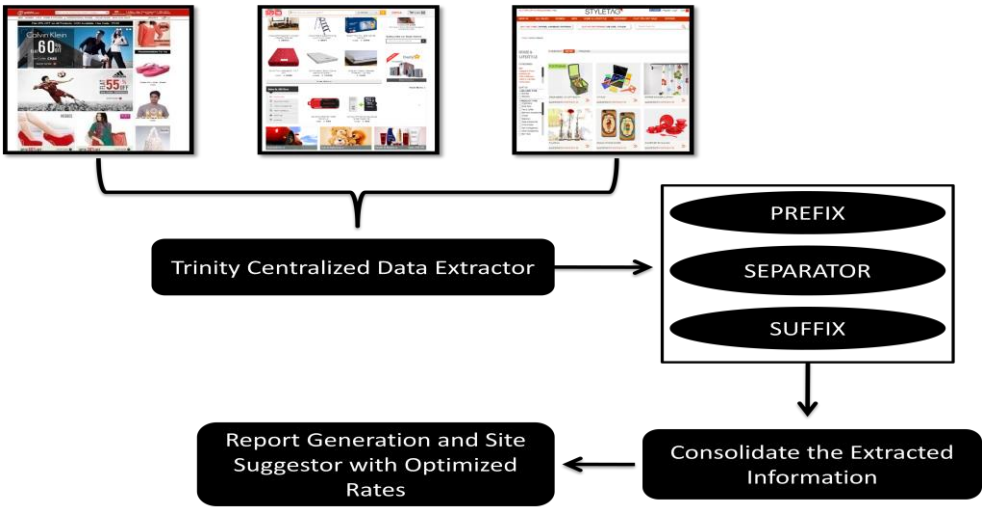

Figure 1. FRAME WORK

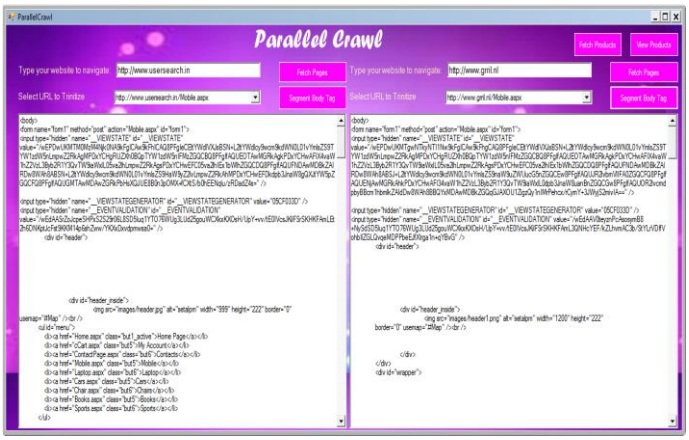

Figure 2. PARALLEL CRAWLING

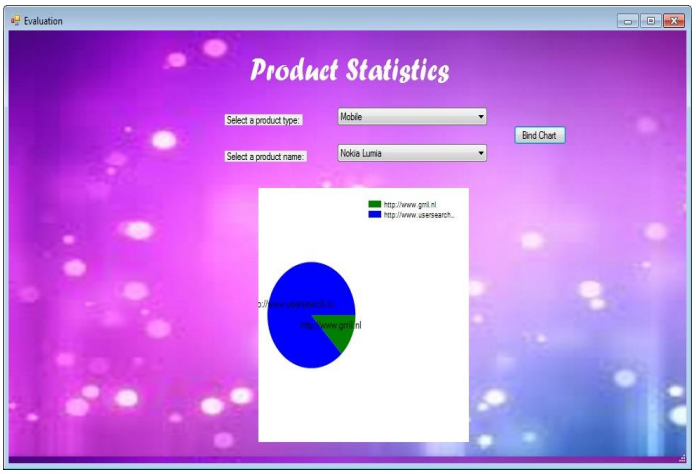

Figure 3. COST COMPARISION CHART

3. Conclusion

Web Mining is the application of data mining techniques to discover interesting usage patterns from the Web data in order to understand and better serve the needs of web-based applications. Online shopping or e-shopping is a form of electronic commerce, which allows consumers to directly buy goods or services from a seller over the internet using a Web browser. The proposed system is able to crawl the multiple website contents and consolidating it to provide data essential for the users**.** It reduces users search time and recommends the best product with low cost.

4. Future Enhancement

Due to the ample shopping websites, the challenging task for the users is to determine which one to prefer. The challenge is also to make sure that they have visited the best website. To pick the best online Web site, the websites should be compared. The comparison involves the cost of the product. This comparison will allow the user to select the best website. In future the proposed system may be modified with more specifications and effective processes.

References

[1] A. Arasu and H. Garcia-Molina, “Extracting structured data from web pages,” in Proc. 2003 ACM SIGMOD, San Diego, CA, USA,pp. 337–348. http://dx.doi.org/10.1145/872757.872799

[2] J. L. Arjona, R. Corchuelo, D. Ruiz, and M. Toro,"From wrapping to knowledge," IEEE Transactions of Knowledge and Data Engineering., vol. 19, no. 2, pp.310–323, Feb. 2007. http://dx.doi.org/10.1109/tkde.2007.31

[3] F. Ashraf, T. Özyer, and R. Alhajj, "Employing clustering techniques for automatic information extraction from HTML documents,"IEEE Trans. Syst. Man Cybern. C, vol. 38, no. 5, pp.660–673, Sept. 2008. http://dx.doi.org/10.1109/tsmcc.2008.923882

[4] M. E. Califf and R. J. Mooney,"Bottom-up relational learning of pattern matching rules for information extraction," *J. Mach. Learn.Res.*, vol. 4, pp. 177–210, May 2003.

[5] A. Carlson and C. Schafer, "Bootstrapping information extraction from semi-structured web pages," in *Proc. ECML/PKDD*, Berlin,Germany, 2008, pp. 195–210. http://dx.doi.org/10.1007/978-3-540-87479-9_31

[6] C. H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A survey of web information extraction systems," *IEEE Transactions of Knowledge and Data Engineering, Vol*. 18, no. 10, pp. 1411–1428, Oct. 2006. http://dx.doi.org/10.1109/tkde.2006.152

[7] 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. http://dx.doi.org/10.1109/mis.2004.71

[8] Hassan A. Sleiman and Rafael Corchuelo (2014), "Trinity: On Using Trinary Trees for Unsupervised Web Data Extraction," *IEEE Transactions of Knowledge and Data Engineering,* vol. 26, no. 6. http://dx.doi.org/10.1109/tkde.2013.161

[9] Hazem Elmeleegy, Jayant Madhavan and Alon Halevy, "Harvesting Relational Tables from Lists on the Web," Proc. of the VLDB Endowment VLDB Endowment Homepage archive Vol. 2 Issue 1, pp.1078-1089, August 2009. http://dx.doi.org/10.14778/1687627.1687749

[10] Valter Crescenzi and Giansalvatore Mecca, "Automatic information extraction from large websites," Journal of the ACM. Volume 51 Issue 5, September 2004, Pages 731-779. http://dx.doi.org/10.1145/1017460.1017462

[11] Yanhong Zhai and Bing Liu, "Web Data Extraction Based on Partial Tree Alignment, " Proc. 2005 Proceedings of the 14th international conference on World Wide Web Pages 76-85. http://dx.doi.org/10.1145/1060745.1060761

Received: April 10, 2015; Published: May 11, 2015