International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 14 (2015) pp 34879-34882

Investigation On The Effect Of Content Extraction In Heterogeneous Web Pages

[1]G. Naveensundar and [2]Dr. A. P. Haran

[1] *Karunya University, Department of CSE, Coimbatore,*
[2]*Park Engineering College, Aeronautical Engineering Department, Coimbatore*
Email: [1]naveensundar@karunya.edu , [2]haran_pct@gmail.com

ABSTRACT:
Web Mining is the application of data mining techniques to automatically discover and extract information from web data. It also uses the data mining techniques to enhance the effectiveness of our interaction with the web. The information available on the web is huge and contains more unnecessary information called as template. Template leads to poor performance of search engine due to the retrieval of non contents for users. Performance can be improved by making the web pages free of template. Furthermore, content extraction is used to identify the main content and remove clutter from the web pages. The hybrid approach of template detection and content extraction helps to remove the templates and extract useful content from web pages. LSH and Min Hash algorithms are used to cluster similar web pages. The proposed approach increases the overall performance (i.e)accuracy of extracted content, reduces the computational time and space.

Keywords:- Cluster, Non-Content Path, Template Detection, Locality Sensitive Hash, MinimumDescription Length

1. INTRODUCTION

In recent years the development of World Wide Web exceeded all expectations. Nowadays there are several billions of pictures, HTML documents, and other multimedia files available via internet and the amount is still rising. But considering the striking variety of the web, retrieving interesting content has become a very daunting task. Web pages in the Websites are constructed in such a way that almost 50% of the data contains templates. This percentage is still increasing as time goes. Templates are a foundation on which actual content is built. From a user point of view, presence of templates is very much useful as they provide uniformity in look and feel of web pages. At the same time, presences of templates in very large amount in web pages compromise the performance of search engines. Hence it is required that the templates should be removed from web pages so that search engines can give good performance in terms of providing the most relevant information in response to user queries. Many of the existing systems were based on the assumption that all the web pages under consideration are built using the same type of template. Such an assumption is not valid in most cases as web pages are built using different types of template structures. Hence this paper is based on the assumption that web pages under consideration are of different types. The structure of templates is different in those pages. LSH and Min hash clustering algorithms are used to cluster similar web pages.

The clustered web pages are fed into the process of content extraction. The content extraction is progressively applied in various fields and applications. Content Extraction is beneficial for visually impaired and blind by identifying the real content within a web page and then increase the font size of the portions of the web page containing contents for better visualization or directly transforming the contents of the web page to speech. The content extraction is used in fields of Natural Language Processing (NLP) and information Retrieval (IR). Where these models derive accurate results based on relevance of contents and the reduction of "standard word error rate (S.Gupta,*et al*, 2003). Most of the NLP based IR applications necessitate dedicated extractors for each of the web domain (S.Gupta,*et al*, 2003) .The generalized content extractors are sufficient and less laborious than hand tailored extractors but is often found less accurate (S.Gupta,*et al*, 2003). A generalized traditional web page contains a title banner, list of links in right or left or both for site navigation and advertisements, a footer containing copyright statements, disclaimers or even sometimes navigational links (T.Weninger,*et al*, 2010). The recent web pages tend to have more cleaner architecture using various layers for visual presentation , real content and interaction (T.V.Raman, 2009) having abandoned the use of old structural tags and adopted an architecture that makes use of the style sheets and div or span tags(T.Weninger,*et al*, 2010). This change in architecture eases the development process but complicates the extraction process hereby reducing the effectiveness of old content extraction systems. The old content extraction systems operate on any varied web pages without considering the type of content the web page represents, which in return yield less accuracy in the content extraction models. This paper proposes methodology to improve accuracy by identifying the content type.

2. RELATED WORK:

Efficient Content extraction Using Hybrid Technique:

This approach is called hybrid because it mainly operates on two models of content extraction one based on statistical features and other on Formatting characteristics. It functions on DOM tree representation of web page calculating the different statistical features associated with the different nodes of tree to measure their importance in providing the

information. The different statistical values like text density and link density for each node are calculated. These values are normalized, so important content should be retained. The calculation is based on the fact that the nodes associated with the content have higher values for the quantity of the text and lower values for quantity of hypertext. In order to achieve an optimal performance on different styled WebPages quantities are normalized with respect to each page. Once the statistically useful nodes are identified, other nodes similar to useful nodes based on formatting characteristics and their position in the page are identified. All of the nodes classified as useful and nodes similar to useful nodes are considered to be the nodes containing real contents.The proposed model identifies the layout of the page by: comparing the quantity of text across each unique node *[ø(i)]* in the DOM tree with the arithmetic mean of the quantity *[Avg ø(T)]* of text across all of the nodes in DOM tree. The deviation *[D(i)]* in text quantity at each node from the arithmetic mean signifies how much contribution of node to the information being rendered to the user. The higher the deviation, the more information is rendered through that nod**e.**

3. ARCHITECTURE FOR TEMPLATE DETECTION:

According to the template detection approach, the paths from the parsed web documents are extracted and the supports of the paths are determined. The Essential paths are the found out based on the following two assumptions

(i) Template path support is higher than the content path support.

(ii) Paths belonging to the template are generally higher than the paths belonging to the content. The process of clustering is then performed using TEXT-MDL algorithm. As a result of clustering, member documents and template paths in the clusters are determined. As a fast approximation of the above method, clustering is then performed using the MinHash concept and the corresponding clusters are determined. To improve the performance, clustering is then performed using the Locality Sensitive Hashing method and finally the performances of all the methods are compared.

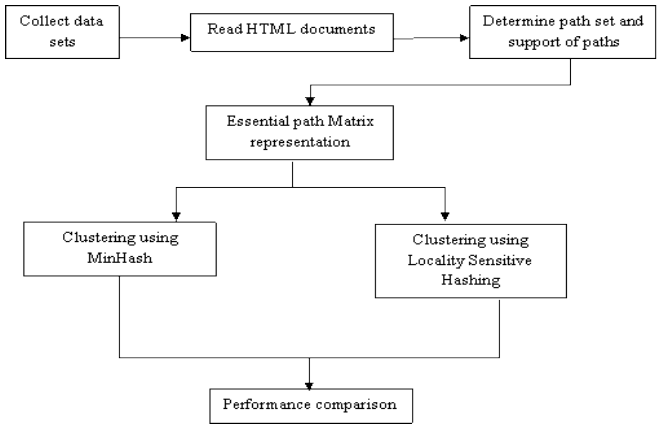

Fig. 1:Template Detection process

4. ARCHITECTURE FOR CONTENT EXTRACTION:

The context extraction process involves extracting the context of the web page to perform the content extraction more accurately. In Fig.2 the architecture of context extraction is shown along with the old system. In this the web page will be given as input to the module of Content genre determination. This module will have the list of top 200 pre classified websites, the new web page will be inspected to classify according to the domain to which it belongs. This can be done by extracting the self-description part of the web page i.e. calculated. The determination of content genre plays useful role while determining the useful node of the web page before determining the useful node in the web pages, it checks the context of the web page and identifies which threshold values to be calculated for the web page.

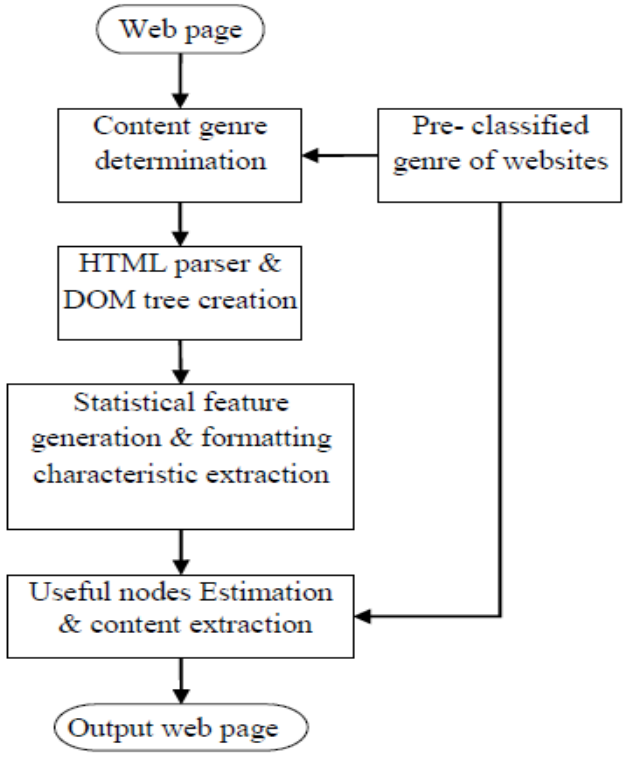

Fig. 2: Content Extraction process

5. RESULTS DISCUSSION:

Sample web pages are taken from ten different websites to calculate statistical features. The text Deviation (D) values obtained are shown in below Fig. 5. Deviation will yield lower values for the nodes which have smaller quantities of text associated with them. The normalized deviation (N) obtained identifies the less informative nodes of DOM tree. The less informative nodes will yield very low values while the nodes which are used in rendering the most of the information have values nearer to 1.

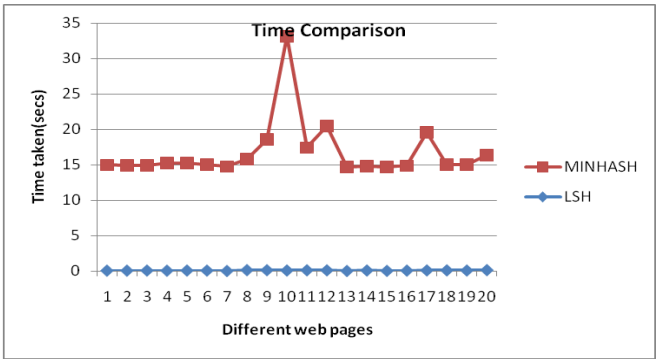

Fig. 3: Time Comparison

Fig. 3 shows the time comparison using LSH and Min Hash. It is clearly shown that the time taken by LSH algorithm is very less when compared to the Min Hash algorithm. Fig. 4 shows that the memory usage is very less in LSH when compared to Min Hash. The method which yields high probability cluster is further fed into the process of Content Extraction.

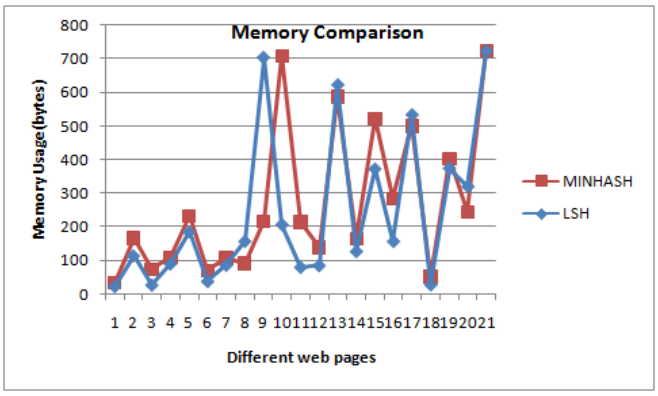

Fig. 4: Memory usage comparison

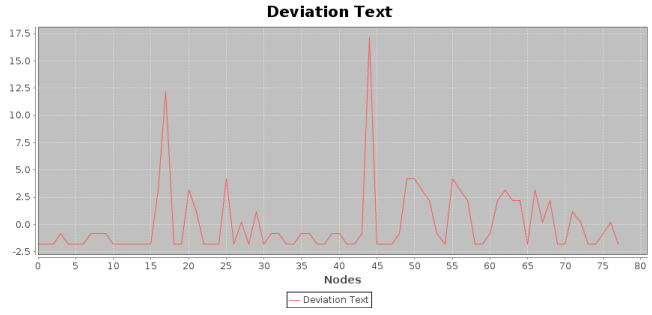

Fig. 5: Deviation in Text

Conclusion:

Most web pages contain huge amount of templates around the body of an article. Although many researches have been done on template detection and content extraction, it is still relatively an emerging field. The proposed research works with Document Object Model tree to perform template detection and Content Extraction, preserving the original data. The techniques employed in this research are quite effective and can efficiently detect templates and extract useful information from web pages.

REFERENCES:

[1] B. Adelberg,"NoDoSE – a tool for semi automatically extracting structured and semistructureddata from text documents", SIGMOD Rec.27 (1998), 283–294.

[2] Z. Bar-Yossef and S.Rajagopalan, "Template Detection via Data Mining and Its Applications", *Proc.11th Int'l Conf.* World Wide Web(WWW), 2002.

[3] K. Vieira, A.S. da Silva, N. Pinto, E.S.de Moura, J.M.B. Cavalcanti and J. Friere, "A Fast and Robust Method for Web Page Template Detection and Removal', *Proc.15th ACM Int'l Conf. Information and Knowledge Mgmt.* (CIKM), 2006.

[4] L.Yi, B. Liu and X. Li, "Eliminating noisy information in Web Pages for Data Mining", *In Proceedings of the International ACM Conference on Knowledge Discovery and Data Mining*, 2003

[5] L. Ma, N. Goharian, A. Chowdhury and M. Chung, "Extracting Unstructured Data from Template Generated Web Documents", *Proc. CIKM*, pp 512-515, 2003.

[6] T. Weninger, T., W.H.Hsu and J. Han, "CETR: content extraction via tag ratios", *Proceedings of the 19th International Conference on World Wide Web*, *ACM*, New York, NY, USA, pp. 971–980, 2010.

[7] A. Arasu and H.Garcia-Molina, "Extracting Structured Data from Web Pages", *Proc.ACM SIGMOD*, 2003.

[8] Liang Chen, Shaozhi Ye, Xing Li, "Template Detection for large scale search engines", *Proc.ACM Symposium*, pp. 1094-1098, 2006.

[9] Jushmerick. N, "Learning To Remove Intsernet Advertisements", *AGENT-99*, 1999.

[10] Yu Wang, Bingxing Fang, Xueqi Cheng, Li Guo and Hongvo Xu, "Incremental Web Page Template Detection", *Proc.17th Int'l Conf. World Wide Web*(WWW), pp. 1247-1248, 2008

[11] M. De Castro Reis,P.B. Golgher, A.S.da Silva and A.H.F. Laender, "Automatic Web News Extraction Using Tree Edit Distance", *Proc.13th Int'l Conf.* World Wide Web(WWW), 2004.

[12] Sandip Debnath, Prasenjit Mitra, C. Lee Giles, "Automatic Extraction of Informative Blocks from Web Pages", Proc. ACM Symposium, pp. 1722-1726, 2005.

[13] S. Zheng, D. Wu, R. Song, J-R. Wen,"Joint Optimization of Wrapper Generation and Template Detection", Proc. ACM SIGKDD, 2007.

[14] V. Crescenzi, P. Merialdo, P. Missier,"Clustering Web Pages Based on Their Structure", Data and Knowledge Eng., vol. 54, 2005, pp. 279-299.

[15] H. Zhao, W. Meng, C. Yu,"Automatic Extraction of Dynamic Record Sections from Search Engine

Result Pages", Proc. 32nd Int'lConf. Very Large Data Bases (VLDB), 2006.

[16] D. Gibson, K. Punera, A. Tomkins,"The Volume and Evolution of Web Page Templates", Proc. 14th Int'l Conf. World Wide Web (WWW), 2005.

[17] A. Kolcz, W. Yih,"Site-Independent Template Block Detection," Proc.KDD, Vol. 4702, 2007, pp. 152-163.

[18] Malcolm Slaney, Michael Casey,"Locality- Sensitive hashing for Finding Nearest Neighbors", IEEE Signal Processing magazine, March 2008.