A Heuristic Approach for Web Content Extraction

Neha Gupta
Department of Computer Applications
Manav Rachna International University
Faridabad, Haryana, India

Dr. Saba Hilal
Director, GNIT – MCA Institute,
GNIT Group of Institutions,
Greater Noida, U. P, India

ABSTRACT
Today internet has made the life of human dependent on it. Almost everything and anything can be searched on net. Web pages usually contain huge amount of information that may not interest the user, as it may not be the part of the main content of the web page. To extract the main content of the web page, data mining techniques need to be implemented. A lot of research has already been done in this field. Current automatic techniques are unsatisfactory as their outputs are not appropriate for the query of the user. In this paper, we are presenting an automatic approach to extract the main content of the web page using tag tree & heuristics to filter the clutter and display the main content. Experimental results have shown that the technique presented in this paper is able to outperform existing techniques dramatically**.**

Keywords
HTML Parser, Tag Tree, Web Content Extraction, Heuristics

1. INTRODUCTION
Researchers have already worked a lot on extracting the web content from the web pages. Various interfaces and algorithms have been developed to achieve the same; some of them are wrapper induction, NET [18], IEPAD [9], Omini [19] etc. Researchers have proposed various methods to extract the data from the web pages and focused on various issues like flat and nested data records, table extraction, visual representation, DOM, inductive learning, instance based learning, wrappers etc. All of these are either related to comparative analysis or Meta querying etc.

Whenever a user query the web using the search engine like Google, Yahoo, AltaVista etc, and the search engine returns thousands of links related to the keyword searched. Now if the first link given by the user has only two lines related to the user query & rest all is uncluttered material then one needs to extract only those two lines and not rest of the things.

The current study focuses only on the core content of the web page i.e. the content related to query asked by the user.
The title of the web page, Pop up ads, Flashy advertisements, menus, unnecessary images and links are not relevant for a user querying the system for educational purposes.

2. RELATED WORK
Huge literature for mining web pages marks various approaches to extract the relevant content from the web page. Initial approaches include manual extraction, wrapper induction [1, 3] etc. Other approaches ([2] [4] [5] [6] [7] [8] [9] [10]) all have some degree of automation. The author of these approaches usually uses HTML tag or machine learning techniques to extract the information.
Embley & Jiang [11] describe a heuristic based automatic method for extraction of objects. The author focuses on domain ontology to achieve a high accuracy but the approach incurs high cost. Lin & Ming [12] propose a method to detect informative clocks in WebPages. However their work is limited due to two assumptions: The coherent content blocks of a web page are known in advance. (2) Similar blocks of different web pages are also known in advance, which is difficult to achieve.
Yossef & Sridhar [13] proposed a frequency based algorithm for detection of templates or patterns. However they are not concerned with the actual content of the web page.
Lee & Ling [14], the author has only focused on structured data whereas the web pages are usually semi structured.
Similarly Kao & Lin [15] enhances the HITS algorithm of [16] by evaluating entropy of anchor text for important links.

3. THE APPROACH USED
The technique used in this paper works on analyzing the content and the structure of the web page .To extract the content from the web page; first of all we pass the web page through an HTML Parser. This HTML Parser is an open source and creates a tag tree representation for the web page. The HTML Parser used in our approach can parse both the HTML web pages and the XML web pages. Since the tag tree developed by the parser is easily editable and can be constructed back into the complete web page, the user gets the original look and feel of the web page. After the tag tree generation, the extraction of objects and separators are extracted using heuristics to identify the content of user's interest.

4. SYSTEM ARCHITECTURE

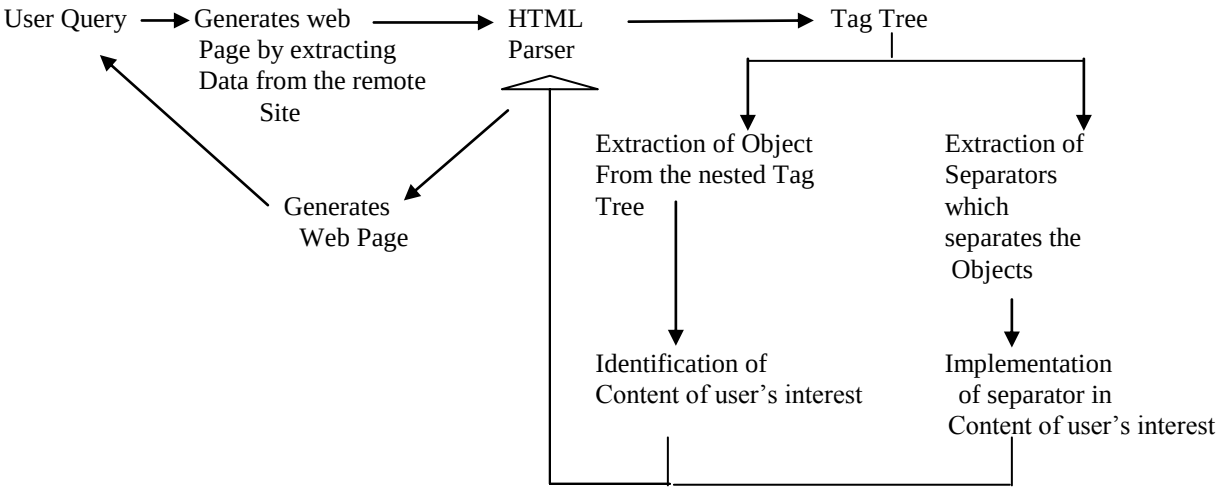

Figure 1. Architecture of the system

4.1 Tag Tree Generations

The web page generated because of the user query is transferred to HTML Parser to generate a hierarchical tag tree with nested object nodes. According to our analysis, group of data records which are of similar type are being placed under one parent node in the tag tree. All the internal nodes of the tag tree are marked as HTML or XML tag nodes and all the leaf nodes of the tag tree are marked as data or content nodes (Content nodes can be any of text, number or MIME data). HTML parser also deals with any type i.e. HTML tag error resiliency.

The HTML parser used will generate independent tag tress for every web page linked to the web site. So we will generate a procedure to integrate all the tag trees into a single tree, having common features of all the web pages. This integrated tree will help us in analyzing the content and structure of the web page.

The tag tree generation can be illustrated with the help of the following figure.

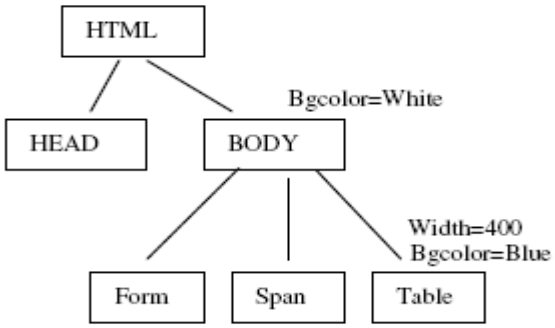

Figure 2.Tag Tree Representation of WebPage1 for XYZ Web site

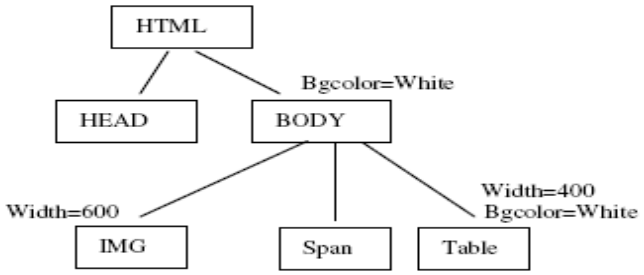

Figure 3.Tag Tree Representation of WebPage2 for XYZ Web site

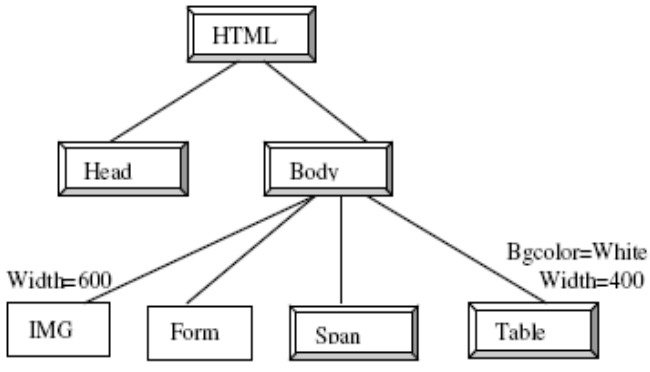

Figure 4. Integrated Tag Tree of the XYZ Web site

From figure 4, we observe that, table and span tag are common and are integrated as such; Form and IMG tags are merged with the common tags as illustrated by the figure. So every node in the integrated tree is a merge node which represents the set of logically comparable blocks in different tag trees. The nodes are merged based upon their CSS features. If the CSS Features match with each other, only then a node can be merged.

4.2 Extraction of objects from the nested tag tree

To extract the nested objects from the tag tree, Depth First Search technique is implemented recursively. While implementing recursive Depth First Search each tag node and content node is examined. The recursive DFS work in two phases. In the first phase, tag nodes are examined and edited if necessary but are not deleted. In the second phase, the tag nodes which refer to unnecessary links, advertisements are deleted based upon the filter settings of the parser.

The tag nodes are deleted so as to keep the content nodes which are of interest to the user, which is also known as primary content region of the web page.

4.3 Extraction of Separators which separates the objects

As soon as we identify the primary content region, the need to separate the primary content from rest of the content arises. To accomplish this task, separators are required. The task of implementing the separators among the objects is quite difficult and needs special attention. To implement the same we will use three heuristics namely pattern repeating, standard deviation and sibling tag heuristics. The details of these heuristics will be under the heading "Implementation of separators in content of user's interest"

4.4 Identification of Content of User's Interest

In this stage, first of all we extract the text data from the web page using separators identified in stage 2. As we have chosen the separators for the objects, now we need to extract the component from the chosen sub tree. Sometime it may happen that we may need to break the objects into 2 pieces according to the need of the separator.

Also in this stage, we refine the contents extracted. This can be achieved by removing the objects that don't meet the minimum standards defined during the object extraction process and are fulfilled by most of the refined objects identified earlier.

4.5 Implementation of separators in content of user's interest

As pointed out earlier, the implementation of separator tags needs to be addressed with the help of three heuristics.

4.5.1 Pattern Repeating (PR):-

The PR heuristics was first proposed by Embley and Jiang [11] in the year 1999. It works by counting the number of paired tags and single tag and rank the separator tag in ascending order based upon the difference among the tags. If the chosen sub tree has no such pair of tags, then PR generates an empty list of separator tags.

4.5.2 *Standard Deviation Heuristics (SDH):-*

The SD heuristics is the most common type of heuristic discussed by many researchers [11] [17]. The SDH measures the distance (By calculating number of characters) between two consecutive occurrences of separator tag and then calculate their SD and rank them in ascending order of their SD.

4.5.3 *Sibling Tag Heuristics (STH):-*

The STH was first discussed by Butler & Liu [17]. The motivation behind STH is to count the pair of tags which have immediate siblings in the minimal tag tree. After counting the pair of tags, they are ranked in descending order according to the number of occurrences of the tag pairs.

5. THE COMBINED APPROACH

The purpose of each heuristic is to identify the separator which can be implemented in content of user's interests.

To evaluate the performance of these heuristics on different web page, we have conducted experiments on 150 web pages. The experimental results in terms of empirical probability distribution for correct tag identification are given in Table 1.

Table 1

Heuristic	Rank 1	2	3	4	5
SD	0.78	0.18	0.10	0.00	0.00
PR	0.73	0.13	0.00	0.00	0.00
SB	0.63	0.17	0.12	0.6	0.03

The Empirical Probability is calculated using additive rule of Probability

P (AUB) =P (A) + P (B) - P (A∩B),

where

P (A) = Probability of Heuristic A on web Page 1,

P (B) = Probability of Heuristic B on web page 1,

and,

P (AUB) = Combined probability of getting correct separator tag on web page 1

To calculate the performance of each heuristic on web page , we use the following formula:

Let 'a' be the number of times a heuristic choose a correct separator tag.

Let 'b' be the no of web pages tested

Let'd' be the no of web sites tested

Then

The success rate given heuristic for each web site 'c'= a / b,

Success rate of each heuristic for total web sites tested = c / d.

Figure 5, 6 and 7, 8 shows the implementation with refined results

Figure 5 – Web Page before Content Extraction

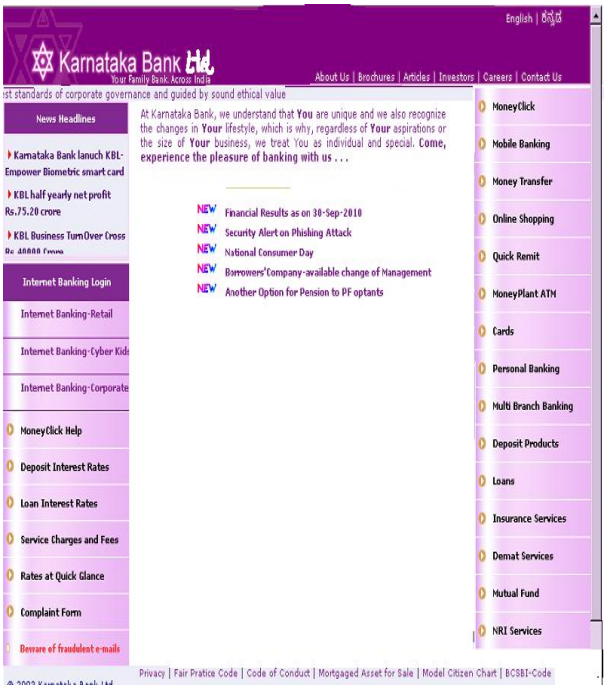

Figure 6 –Web Page after Content Extraction

Figure 7- Web Page before Content Extraction

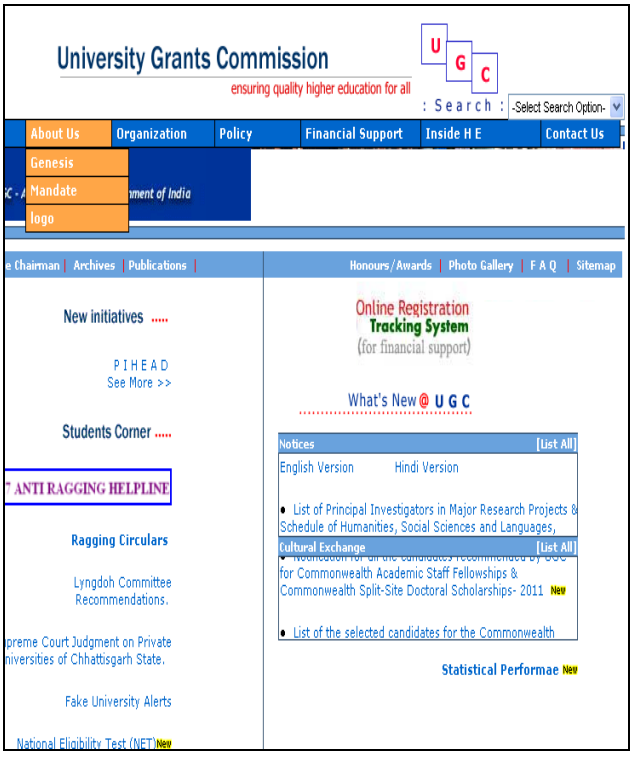

Figure 8 – Web Page after Content Extraction

6. CONCLUSION

In this paper a heuristic based approach is presented to extract the content of user's interest from the web page. The technique that is implemented is simple and quite effective. Experimental results have shown that the web page generated after implementation of heuristic is better and give effective results. The irrelevant content like links, advertisements etc are filtered and the content of user's interest is displayed.

7. REFERENCES

[1] P. Atzeni , G. Mecca, " Cut & Paste" , Proceedings of 16th ACM SIGMOD Symposium on Principles of database systems, 1997

[2] C. Chang and S. Lui, " IEPAD: Information extraction based on pattern discovery" , In Proc. of 2001 Intl. World Wide Web Conf., pages 681–688, 2001

[3] J. Hammer, H. Garcia-Molina et al , " Extracting semi-structured data from the web" , Proceedings of workshop on management of Semi=-Structured Data, Pages 18-25, 1997

[4] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, "A. Rajaraman, Y. Sagiv, J. D. Ullman, and J. Widom. The TSIMMIS project: Integration of heterogenous information sources", Journal of Intelligent Information Systems, 8(2):117–132, 1997.

[5] M. Garofalokis, A. Gionis, R. Rastogi, S. Seshadr, and K. Shim, " XTRACT: A system for extracting document type descriptors from XML documents", In Proc. of the 2000 ACM SIGMOD Intl. Conf. on Management of Data, pages 165–176, 2000.

[6] E. M. Gold, "Language identification in the limit. Information and Control", 10(5):447–474, 1967.

[7] S. Grumbach and G. Mecca, " In search of the lost schema" , In Proc. of 1999 Intl. Conf. of Database Theory, pages 314–331, 1999.

[8] J. Hammer, H. Garcia-Molina, J. Cho, A. Crespo,R. Aranha, "Extracting semi structure information from the

[9] Web", In Proceedings of the Workshop on Management of Semi structured Data, 1997.

[10] Chang, C-H., Lui, S-L, "IEPAD: Information Extraction based on pattern discovery", ACM Digital Library WWW-01, pp 681-688, 2001

[11] A. Laender, B. Ribeiro-Neto et.al, " A brief survey of Web Data Extraction tools" , Sigmod Record, 31(2),2002

[12] D.W. Embley, Y. Jiang, " Record Boundary Discovery in Web Documents", In Proceeding of the 1999 ACM SIGMOD, Philadelphia, USA, June 1999.

[13] Shian-Hua Lin, Jan-Ming Ho, "Discovering informative content blocks from Web documents", SIGKDD-2002, 2002.

[14] Ziv Bar-Yossef, Sridhar Rajagopalan, "Template detection via data mining and its applications" , WWW-2002, 2002

[15] Mong Li Lee, Tok Wang Ling, Wai Lup Low, "Intelliclean: A knowledge-based intelligent data cleaner", SIGKDD-2000, 2000.

[16] Hung-Yu Kao, Ming-Syan Chen Shian-Hua Lin, and Jan-Ming Ho, "Entropy-Based Link Analysis for Mining Web Informative Structures", CIKM-2002, 2002.

[17] Jon M, Kleinberg, "Authoritative sources in a hyperlinked environment." In Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 1998

[18] D. Buttler, L. Liu, C.Pu, "Omini: An object mining and extraction system for the web", Technical Report, Sept 2000, Georgia Tech, College of Computing.

[19] Liu,B, Zhai, Y., "NET- "A System for extracting Web Data From Flat and Nested Data Records", WISE-05 (Proceeding of 6th International Conference on Web Information System Engineering), 2005

[20] Buttler, D. Ling Liu Pu, C., "A fully automated object extraction system for the World Wide Web", IEEE explore ICDCS 01, pp 361-370, 2001

SKETCH OF AUTHORS BIOGRAPHICAL

Prof. Saba Hilal is currently working as Director in GNIT – MCA Institute, Gr. Noida. She has done Ph.D (Area: Web Content Mining) from Jamia Millia Islamia, New Delhi, and has more than 16 years of working experience in industry, education, research and training. She is actively involved in research guidance/ research projects/ research collaborations with Institutes/ Industries. She has more than 65 publications/ presentations and her work is listed in DBLP, IEEE Explore, Highbeam, Bibsonomy, Booksie, onestopwriteshop, writers-network etc. She has authored a number of books and has participated in various International and National Conferences and Educational Tours to USA and China. More details about her can be found at http://sabahilal.blogspot.com/

Ms. Neha Gupta is presently working with Manav Rachna International University, Faridabad as Assistant Professor and is Pursuing PhD from MRIU. She has completed MCA from MDU, 2006 and Phil from CDLU, 2009. She has 6 years of work experience in education and industry. She is faculty coordinator and has received many appreciation letters for excellent teaching.