DOI: 10.21917/ijsc.2014.0099

WEB CONTENT EXTRACTION USING HYBRID APPROACH

K. Nethra[1], J. Anitha[2] and G. Thilagavathi[3]
[1, 3]*Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, India*
E-mail: [1]nethrakanagaraj@gmail.com, [3]thilaga.apr@gmail.com
[2]*Department of Information Technology, Sri Ramakrishna Engineering College, India*
E-mail: anitha.j@srec.ac.in

Abstract
The World Wide Web has rich source of voluminous and heterogeneous information which continues to expand in size and complexity. Many Web pages are unstructured and semi-structured, so it consists of noisy information like advertisement, links, headers, footers etc. This noisy information makes extraction of Web content tedious. Many techniques that were proposed for Web content extraction are based on automatic extraction and hand crafted rule generation. Automatic extraction technique is done through Web page segmentation, but it increases the time complexity. Hand crafted rule generation uses string manipulation function for rule generation, but generating those rules is very difficult. A hybrid approach is proposed to extract main content from Web pages. A HTML Web page is converted to DOM tree and features are extracted and with the extracted features, rules are generated. Decision tree classification and Naïve Bayes classification are machine learning methods used for rules generation. By using the rules, noisy part in the Web page is discarded and informative content in the Web page is extracted. The performance of both decision tree classification and Naïve Bayes classification are measured with metrics like precision, recall, F-measure and accuracy.

Keywords:
Web Mining, Web Content Extraction, Decision Tree Learning, Naïve Bayes Classification, DOM Tree

1. INTRODUCTION

The World Wide Web has grown explicitly which provides access to all people at any place and at any time. It facilitates any one to upload or download relevant data and the valuable content in the Web site can be used in all fields [15]. The data in the Web are unstructured and semi-structured, lots of insignificant and irrelevant document are obtained as a result after navigating several links. Hence data mining cannot be applied directly [20]. For effective retrieval of Web information, Web mining is used.

The application of data mining techniques to automatically discover and to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of Web sites, etc, is called Web mining [9]. Some of the data mining techniques applied in Web mining are association rule mining, clustering, classification, frequent item set. Some of the sub tasks of Web mining are finding of relevant resource, selection of information and preprocessing, generalization and analysis [12].

Web content mining is used for extracting useful information from Web pages. Web page content can be structured, unstructured and semi-structured. Structured Web page data's are easy to extract when compared with unstructured and semi-structured data's [4].

Web Content Extractor normally extracts a whole Web page including links, header, footer, main content and advertisement [10]. During the extraction unwanted data like links, header, footer and advertisement are treated as noisy information. To eliminate the noisy information and extract the useful information is a challenging problem [11]. Many techniques were proposed for eliminating noisy information.

In this paper, we focus on hybrid techniques on Web content extraction and Section 1 gives a brief introduction about Web content mining. The rest of the paper is organized as follows. Section 2 introduces Web content mining approaches which are used in Web data extraction. Section 3 explains the hybrid approach. Section 4 describes the results and discussion. Section 5 gives the conclusion is made.

2. RELATED WORK

Existing Web Content Extraction techniques are grouped into two major categories (i) Automatic Extraction, (ii) Hand-crafted rules generation.

2.1 AUTOMATIC EXTRACTION

Automatic Extraction is the process of extracting the Web page content automatically using tools and techniques [1]. Web page segmentation can be done based on three approaches and they are DOM-based segmentation, location-based segmentation and visual-based segmentation.

Milos kovacevic, Michelangelo Diligenti, Marco Gori, Marco Maggini, Veljko Milutinovic [8] proposed a method for extracting and processing information from Web pages. Visual information from an HTML source is extracted using M-Tree which consists of HTML parser for building a tree. Naive Bayes classifier is used for Web page classification. Based on the probability and score assignment of Web pages, documents are classified.

Christian Kohlschutter, Peter Fankhauser, Wolfgang Nejdl [7] proposed approach for boilerplate detection using shallow Text features which is theoretically ground by stochastic text generation process from Quantitative Linguistics. Text Content of a Web page is grouped into two classes' long text and short text. In systematical analysis words in the short text is removed. Boilerplate detection strategies on four representative multi-domains corpora are evaluated.

Yu Chen, Wei-Ying Ma, Hong-Jiang Zhang [3] proposed a method for facilitating the browser of a large Web page on a small screen. A Web page is adapted and converted into a two level hierarchical organization with a thumbnail page at the top level for providing a global view and index to a set of sub-pages at the bottom level for detail viewing. A page analysis algorithm

is used to extract the semantic structure of an existing Web page and a page splitting scheme to partition the Web page into smaller and logically related content blocks.

Marco Baroni, Francis Chantree, Adam kilgariff, Serge Sharoff [2] proposed CleanEval as a shared task for cleaning arbitrary Web pages. Initially data preparation is done by data selection and annotation is guided with instruction like removal of HTML / Java code and "boilerplate" and adding a basic encoding of the structure of the page using minimal set of symbols to mark the beginning of headers, paragraphs and list elements. Finally Scoring need to measure the similarity between two differently cleaned versions of a file.

Christian Kohlschutter [6] proposed a densitometric analysis of Web Template Content. Web page's noisy data is not useful for classification. So Web content of several large was subjected to a quantitative analysis. By deriving a densitometric text model based upon techniques from the field of quantitative Linguistics.

John Gibson, Ben Wellner, Susan Lubar [5] proposed a method to identify target content of a Web page. Problem of identifying content is solved by sequence labeling method and boundary detection method. Some of the models employed for sequence labeling are Conditional Random Fields (CRF), Maximum Entropy Classifiers (MaxEnt) and Maximum Entropy Markov models (MEMM). Web pages are divided into blocks and features are selected and machine learning technique is applied.

ErdincUzun, TarikYerlikaya, Meltem Kurt [19] developed a SET parser which is utilized regular expressions and string functions for extracting this tags. DOM is used mostly but it is time and memory consuming. This SET method provides less parsing time and memory than use of DOM.

2.2 HAND-CRAFTED RULES

Hand crafted rule generation uses string manipulation function for rule generation. Hand-crafted rules are impractical for more than a couple of data source.

3. A HYBRID APPROACH

Eliminating irrelevant Information and extracting Informative content uses automatic extraction techniques and hand-crafted rule generation. Automatic extraction technique uses machine learning methods while implementing these techniques it increases the time complexity of the extraction process. Extraction through hand crafted rules is an efficient technique but preparing the rules is cumbersome. A hybrid approach is proposed which consists of both automatic extraction and rule generation techniques.

The approach involves generation of automatic rules instead of manual hand-crafted rule insertion. The rules generated are used to infer informative content from simple HTML pages. Initially DOM tree is constructed to demonstrate a visual content of the Web page with richer features. Feature extraction is applied between div and td tags. Machine learning methods like Decision tree classification and Naïve Bayes Classification are applied to generate the rules and create a well formed document. Rules generated are used for extracting the informative content from the Web pages.

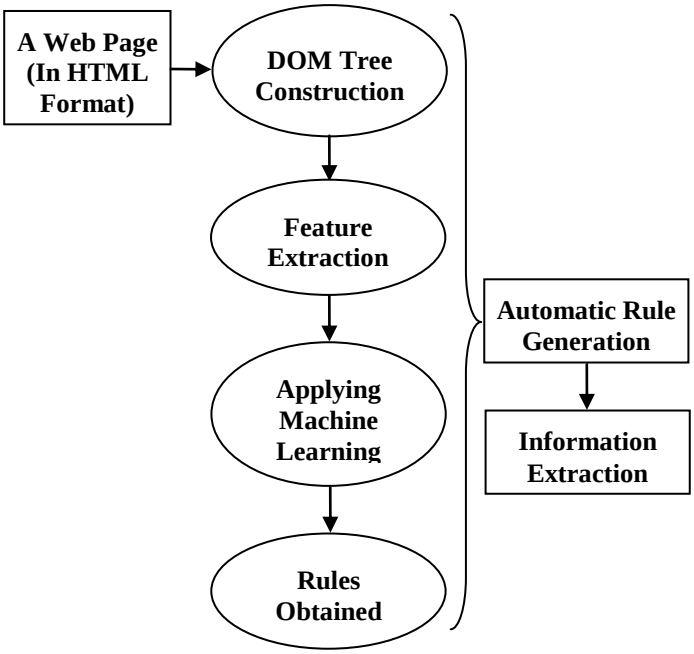

Fig.1. Architecture of a hybrid approach

To automate pattern extraction and maintain efficiency, in hybrid approach, patterns are first obtained as rules by applying machine learning (ML) techniques. Finally, those extracted rules are used for extracting the informative content from Web pages without using ML inference.

3.1 AUTOMATIC RULE GENERATION

A Web page which is designed using a HTML is converted to DOM tree. Features like word frequency, Link frequency, Density in HTML, Average and Ratio of word frequency are calculated and extracted. Based on features, Machine learning technique like Decision Tree classification and Naïve Bayes classification are applied and rules are generated.

In decision Tree classification, C4.5 algorithm is applied .In C4.5 algorithm Entropy and Information gain is calculated for all features. With the help of information gain decision tree is build well formed document. In Naïve Bayes classification, general classification technique is applied to obtain well formed document. Well formed document obtained has rules in XML format.

3.1.1 DOM tree Construction:

HTML is a simple, easy and effective markup language used for developing the Web sites.HTML consists of several tag sets to visualize the content. To demonstrate the visual content richer features of a Web page, a hierarchy called DOM is used. DOM (Document Object model) is a cross platform and language independent convention used for representing HTML documents. With DOM tree, missing tags can be easily identified.

An HTML tag is generally formed with HEAD and BODY tag. HEAD tag consists of necessary information about the Website, and the BODY tag is used to visualize the content. DIV and TD tags are also known as block tags and used to separate the Website into several blocks, and they are referred as block

DOI: 10.21917/ijsc.2014.0099

markers. Block tags have descriptive parts like ID and CLASS which are used to produce the rules.

3.1.2 Extracting Feature:

Block Detection depends on using proper features for extraction process. A parent DIV tag can have both uninformative and informative DIV children. Uninformative DIV children may create noise in the statistics derived from the parent DIV tag, so that the extracted features may not contain appropriate link and word frequency ratios. Most child DIV/TD tag is extracted to form a non-nested structure from the parent DIV/TD tags. Some of the features [18] extracted are,

- Frequency of word is the count of terms within the tags
- Density within HTML is the ratio of the count of terms within tags to the count of all terms within the HTML document
- Frequency of link is the number of A HREF links within tags
- Frequency of words inside the links is the number of terms inside A HREF links placed tags
- Average frequency of words inside links is the ratio of the count of terms inside A HREF links inside tags to the count of links.
- Ratio of frequency of words inside links to all words is the ratio of the number of terms inside A HREF links placed inside tags to all of the count of terms inside tags.

3.1.3 Applying Machine learning Methods:

Machine learning is used to acquire knowledge automatically from existing data. Machine learning is a process by which a system improves its performance. Two Machine learning technique's like decision tree classification and Naïve Bayes classification is used to extract rules.

3.1.3.1 Decision Tree classification:

A decision tree [16] is used for decision making purpose. Decision tree has root and branch node. From the root node, users split each node recursively based on decision tree learning algorithm. The final result of decision tree consists of branches and each branch represents a possible scenario of decision and its consequences. C4.5 algorithm [17] is applied as decision tree classification.

Pseudo code of C4.5 algorithm

- Initially check for base cases
- For each attribute a
  - By splitting the attribute a, the normalized information gain is found.
  - Let a_best be the attribute with the highest normalized information gain.
  - Then create a decision node which splits on a_best
  - Recurse on the sub lists obtained by splitting on a_best and add those nodes as children of node.

Entropy Formula

Entropy is one kind of measurement procedure in information theory,

$$Entropy(S) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right) \quad (1)$$

where, p is positive value and n is a negative value.

Information gain Formula

An information gain is used to find best attribute. Gain measure gives how a given attribute separates the training example into targeted classes. Rules are selected based on gain ration,

Gain = Current feature of Entropy – Sum of all Entropy (2)

3.1.3.2 Naïve Bayes classification:

A Naive Bayes classification [13] is a probabilistic classification method which is based on Bayes theorem with strong independence assumptions. A Naive Bayes Classification [14] is a program which predicts a class value given a set of set of attributes.

Pseudo code of Naïve Bayes classification

For each known class value,

- Calculate probabilities for each attribute, conditional on the class value.
- Use the product rule to obtain a joint conditional probability for the attributes.
- Use Bayes rule to derive conditional probabilities for the class variable.

Once this has been done for all class values, output the class with the highest probability.

3.1.4 Obtaining rules from ML results:

A well formed document is created in XML format. The document consists of main tag with informative content. By using the rule which is obtained from well-formed document can be used to extract the main content from Web page.

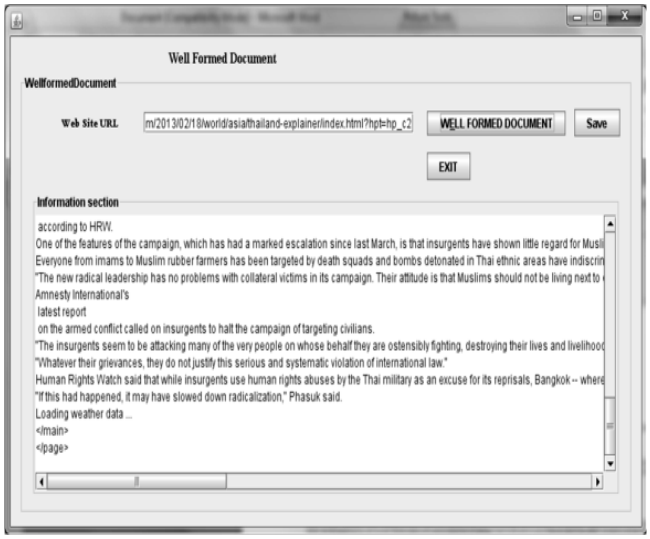

Fig.2. Well formed document

3.2 EFFICIENT WEB CONTENT EXTRACTION

Efficient Web Content Extraction is used to extract the Web content. Web page content extraction is based on simple string manipulation functions like search and substring. The count of

start and end tags are calculated to verify whether the current rule obtained is proper for extracting the content. If the count of start and end tags are equal, the content can be easily extracted from substring function.

4. RESULTS AND DISCUSSION

A Web page which is given as an input is converted to DOM tree. Features are extracted and given as an input for machine learning methods like Naïve Bayes classification and C4.5 decision tree classification. From that rules are generated and using the rules informative content of the Web page is extracted.

Performance of Naïve Bayes classification and C4.5 decision tree classification method is obtained by calculating the metrics. Metrics like precision, recall, F-measure and accuracy are used. The accuracy metric is used to measure the percentage of correct predictions for the overall data.

Precision finds the fraction of records which actually turns out to be positive in the group where the classifier has declared as a positive class. Recall finds the fraction of correct instances among all instances that actually belong to a relevant subset. A measure which combines precision and recall is F-measure, which can be also known as the harmonic mean of precision and recall.

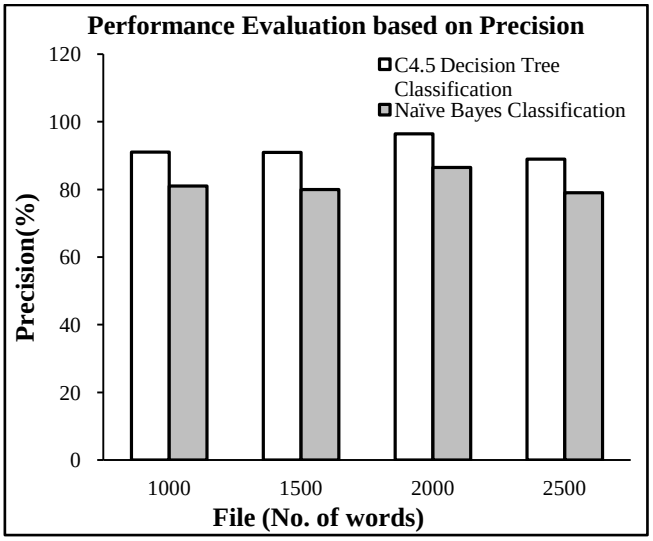

Fig.3. Performance comparison based on precision

The Fig.3, gives the comparison of C4.5 decision tree classification and Naïve Bayes classification based on the metric precision. When numbers of words in a HTML file increases, precision of C4.5 decision tree classification is high when compared with Naïve Bayes classification. C4.5 decision tree classification achieves 89% precision whereas Naïve Bayes classification achieves only 81% precision.

The Fig.4, gives the comparison of C4.5 decision tree classification and Naïve Bayes classification based on the metric recall. When numbers of words in a HTML file increases, recall of C4.5 decision tree classification is high when compared with Naïve Bayes classification. C4.5 decision tree classification achieves 81% recall whereas Naïve Bayes classification achieves only 76% recall.

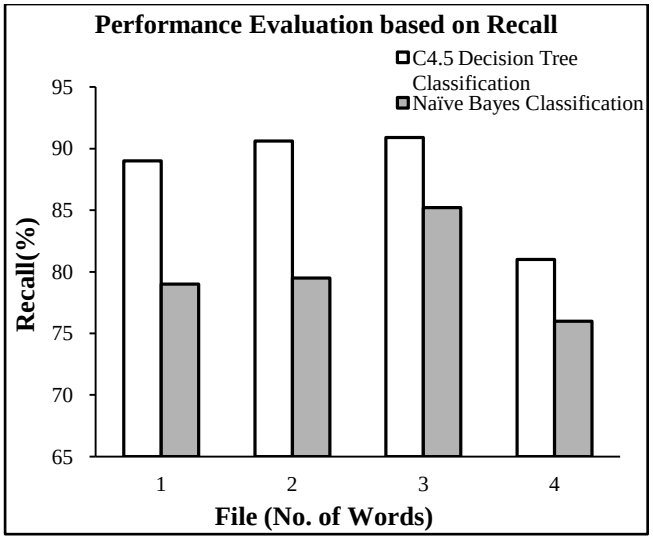

Fig.4. Performance comparison based on recall

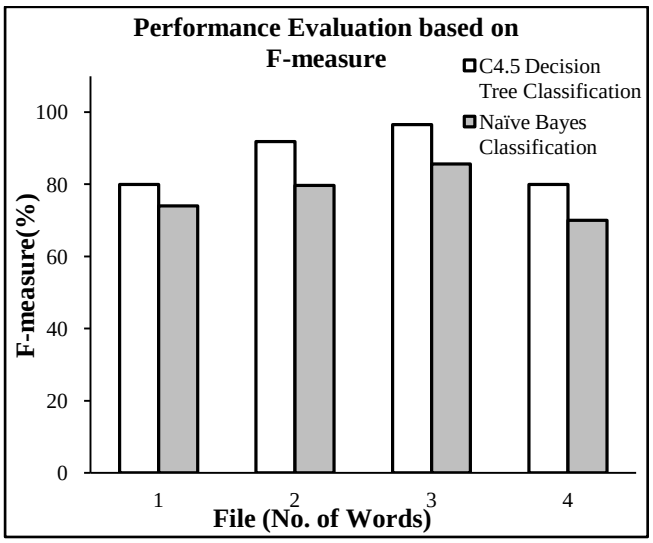

Fig.5. Performance comparison based on F-measure

The Fig.5, gives the comparison of C4.5 decision tree classification and Naïve Bayes classification based on the metric f-measure. When numbers of words in a HTML file increases, f-measure of C4.5 decision tree classification is high when compared with Naïve Bayes classification. C4.5 decision tree classification achieves 80% f-measure whereas Naïve Bayes classification achieves only 70% f-measure.

The Fig.6, gives the comparison of C4.5 decision tree classification and Naïve Bayes classification based on the metrics accuracy. When numbers of words in a HTML file increases, accuracy of C4.5 decision tree classification is high when compared with Naïve Bayes classification. C4.5 decision tree classification achieves 55% accuracy whereas Naïve bayes classification achieves only 53% accuracy.

DOI: 10.21917/ijsc.2014.0099

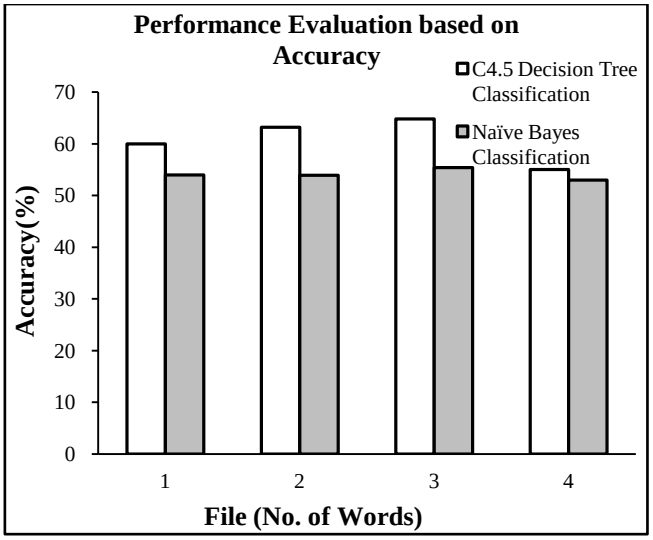

Fig.6. Performance comparison based on Accuracy

5. CONCLUSION AND FUTURE WORK

Informative content from the Web page is extracted and noisy data's like links, advertisement, header, footer etc are avoided. A hybrid approach which is a combination of automatic content extraction and rule generation is applied to extract informative content. A Web page is converted to DOM tree and features are extracted. Extracted features are used as an input for machine learning methods like decision tree classification and Naïve Bayes classification. Rules are generated as the result of machine learning methods and well formed document is obtained. Informative content of Web page can be extracted by using the generated rules. Performances of machine learning methods are compared with metrics like precision, recall, f-measure and accuracy. As a result C4.5 decision tree classification algorithm performs better than the Naïve Bayes classification.

This method can be used in Web crawler for automatic efficient Web content extraction and fuzzy techniques can be employed for rule generation.

REFERENCES

[1] S. Baluja, "Browsing on smalls screens: Recasting Web-page segmentation in to an efficient machine learning framework", *Proceedings of the 15th International Conference on World Wide Web*, pp. 33–42, 2006.

[2] M. Baroni, F. Chantree, A. Kilgarri, S. Sharoff, "Cleaneval: A competition for cleaning Web pages", *Proceedings of the sixth International Conference on Language Resources and Evaluation*, 2008.

[3] Y. Chen, W. Y. Ma and H. J. Zhang, "Detecting Web page structure for adaptive viewing on small form factor devices", *Proceedings of the 12th International Conference on World Wide Web*, pp. 225–233, 2003.

[4] S. Debnath, P. Mitra, N. Pal and C. L. Giles, "Automatic identification of informative sections of Web pages", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 9, pp. 1233–1246, 2005.

[5] J. Gibson, B. Wellner and S. Lubar, "Adaptive Web-page content identification", *Proceedings of the 9th annual ACM International workshop on Web Information and Data Management*, pp. 105–112, 2007.

[6] C. Kohlschutter, "A densitometric analysis of Web template content", Proceedings of the 18th International Conference on World Wide Web, pp. 1165-1166, 2009.

[7] C. Kohlschutter, P. Fankhauser and W. Nejdl, "Boiler plate detection using shallow text features", *Proceedings of the third ACM International Conference on Web search and data mining*, pp. 441–450, 2010.

[8] M. Kovacevic, M. Diligenti, M. Gori and V. Milutinovic, "Recognition of common areas in a Web page using visual information: A possible application in a page classification", *Proceedings of the IEEE International Conference on Data Mining*, pp. 250-257, 2002.

[9] S. Mahesha, M. S. Shashidhara and M. Giri, "An Efficient web content extraction using mining techniques", *International Journal of Computer Science and Management Research*, Vol. 1, No. 4, pp. 872-875, 2012.

[10] Nikolaos Pappas, Georgios Katsimpras and Efstathios Stamatatos, "Extracting Informative Textual Parts from Web Pages Containing User-Generated Content", *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, 2012

[11] K. C. Srikantaiah, M. Suraj, K. R. Venugopal, S. S. Iyengar and L. M. Patnaik, "Similarity Based Web Data Extraction and Integration System for Web Content Mining", *Proceedings of International Conference on Communication Network and Computing*, pp. 269–274, 2012.

[12] Sachin Bojewar, Varsha Bhosale, Shuveta Chanchlani, "Data Extraction from dynamic web pages based on visual features", *International Journal of Advanced Engineering Research and Studies*, Vol. 1, No. 1, pp. 91-94, 2012.

[13] Erdinc Uzunc, Hayri Volkan Agun and Tarik Yerlikaya, "A hybrid approach for extracting informative content from Web pages", *Information Processing and Management*, Vol. 49, No. 4, pp. 928-944, 2013.

[14] Erdinc Uzun, Tarik Yerlikaya and Meltem Kurt, "A lightweight parser for extracting useful contents from Web pages", *Proceedings of 2nd International Symposium on Computing in Science & Engineering*, pp. 67-73, 2011.

[15] Tarik Yerlikaya and Erdinc Uzun, "An intelligent browser viewed main content in web pages", *İnternet Sayfalarında Asıl İçeriği Gösterebilen Akıllı Bir Tarayıcı*, pp. 53–57, 2010.

[16] http://en.wikipedia.org/wiki/Decision_tree_learning

[17] http://en.wikipedia.org/wiki/C4.5_algorithm

[18] http://en.wikipedia.org/wiki/Naive_Bayes_classifier

[19] http://www.sussex.ac.uk/Users/christ/crs/ml/lec02b.html

[20] http://www.win.tue.nl/~mpechen/projects/pdfs/Louvan2009.pdf.