International Journal of Advanced Research in Engineering Technology & Sciences ISSN: 2394-2819
May-2016 Volume 3, Issue-5
Email:editor@ijarets.org www.ijarets.org

A SMART WAY FOR CRAWLING INFORMATIVE WEB CONTENT BLOCKS USING DOM TREE METHOD

Gowri. V
Student
MCA Department
MREC, Thathanur, Tamil Nadu

Siva Prakash. R
Assistant Professor
MREC, Thathanur
Tamil Nadu

ABSTRACT
Normally, lots of web pages are available on the web. It discovers the useful knowledge to the researchers. The exact information should retrieval from the web. It is now the great challenge for the researchers to use new methodologies for web mining. The search engines are used crawler system for gather information from the web. Search engine takes into account only the informative content for indexing. A crawler is one of the basic needs of information extraction. Web mining has 3 sub categories, there are, Web content mining, Web Usage Mining and Web Structural Mining. Web content mining is the process of identify the precise data from text, images, audio or video data already available on the web. Web content mining basically used for information retrieval and natural language processing. Web pages commonly have two parts informative and non informative content. The researchers want informative content only. Large amount of noise appear in the non informative content. Each web page has several blocks. Some blocks showing useful content, and some blocks showing noisy content, such as, banners, advertisements and copyright etc... The main motive in a proposed technique is to remove the noisy content blocks and extraction the informative content blocks from web page is based on the idea of web page segmentation. Here, web page is segment into n blocks and the block calculated for each block. Importance blocks are considered as important blocks and the remaining blocks are eliminated as noisy blocks. The proposed approach saves important space and time and shows useful information without noisy blocks.

KEYWORDS
Search engine, Information extraction, web mining, web content mining, web page segmentation, repetition detection, informative blocks, non-informative blocks and noise.

INTRODUCTION
Web contents are segmented into several blocks that blocks may be a multimedia data, structured i.e. XML documents, semi-structured i.e. HTML documents and unstructured data i.e. plain text [1]. These are all providing useful information to the users. Basically useful information combined with noisy content. Both contents are discovering on the web at the same time. The noisy content take some place on the web page so, the informative content has limited space. Web mining takes such as web page clustering, classification/categorization, information retrieval and information extraction.
Various sub tasks also available in web mining there are, Resource finding, Information Selection and Pre-processing Generalization, Analysis and Visualization. The unwanted blocks are useful for limited users only, it's not useful are all researchers, it's essential for the web site owners. But slow down automated information gathering. So it makes irritated to surf the users. There are at least four different known categories of noise patterns within web pages of any web sites including banners with links including search panels, advertisements, navigational panel (directory list) and copy right and privacy notice in each web site. Many web pages contain these four noise categories together but most of the noise patterns are structured by using sectioning tags and sectioning separating tags and interactive tags. Moreover, anchor tags are most commonly used to link another web page or another web site. Internet used only on pc in past days nowadays internet is possible to using mobiles and also PC. So when occur

the noisy content in the web page it makes the slow process when we using mobile devices because, the web page displayed in a small size for using small devices, if noisy blocks occur, takes more space. So the informative content should not possible to display in mobile devices. All disturbances basically occur by non-informative content. The performance for improving the traditional information retrieval, it is vital to differentiate valuable information from noisy content that may mislead users notice within a solitary web page. Removing the non-informative content is thus of great importance.

In this paper presents a technique of information extraction for extract informative content blocks. Section II describes the various techniques to extract informative content. Section III describes the proposed technique. Section IV about the result and finally section V presents future work and conclusion.

LITERATURE SURVEY

Main problems occur in the web is that the informative content on web site is often mixed with non-informative content. Search engine uses the crawler system for Parsing and indexing the vast amount of information on the web is serious challenge.

Elimination of noisy and irrelevant contents from web pages has many applications, including web page classification, clustering, web featuring, proper indexing of search engines, efficient focused crawlers, cell phones and PDA browsing, speech rendering for the visually impaired, improving the quality of search results and text summarization. Thus cleaning web pages for web information extraction becomes crucial for improving the performance of information retrieval. We analyze to remove various noise patterns in web pages instead of extracting relevant informative content blocks from web pages. In this work, we focus on identifying and removing noises in web pages to improve the performance of web content mining considering the fact that local noises in web pages can seriously harm the accuracy of mining. Web page segmentation is one of the approaches to identify the informative content blocks. This section provides a summary of the techniques used to identify the informative content blocks within a web page. Two methods are used in this section. One is template-based method and another one is tag-based method.

The template-based method is focus on measuring the structural patterns among the DOM trees of web pages. Templates play an important role in web pages. Two types of templates are occurring in all web pages. Useful templates and another one are irrelevant templates. So the template-based method builds a useful template with extracting rules organized by regular expressions [8]-[10]. It collects web page from achieving site and generates ordinary expression rules in order to extract content blocks through analyzing the common zone. That it is impossible to segment information contained in a tree node because the template-based method recognizes only the structural information, not the content, of DOM tree nodes. Classifier is the method proposed by Mr. Chakrabarti. Many web pages to train a classifier and build a template, causing the segmentation to be restricted. Visual information in the web page is one of the approaches to overcome these difficulties.

Next a vision-based method it utilizes visual clues in a web page. This method proposed by Chen. This method considers visual information such as <HR>,
 tags [12] [13]. A three stage algorithm is proposed with phases *featuring, modeling and pruning*. It is proposed by Mr. Das. It combining a different term weighting approach (for main content, URL, heading, title, anchor text, information in the meta-tags etc.). It is to find how to identify noisy blocks or irrelevant blocks from an input record, the web page to be processed, with a reduced complexity and increased efficiency of web page [4]. Vision-based page segmentation algorithm considering vision information and rules to identify blocks it is proposed by Yang [14]. This algorithm segments page in three steps: (1) Extracting visual blocks, (2) Detecting separators between extracted blocks and (3) Detecting content structure based on results of previous two steps.

Next we using tag-based method it contains predefined content tags that tags provides useful information and finds web content blocks by measuring the distance between these tags [1], [4], [12], [13], [14], [15], [16], [17]. In particular, Lan yi proposed *compressed structure tree* to compress or to merge a set of Web pages from a Web site. Then propose a measure that is based on information theory to evaluate each element node in the compressed structure tree to determine the importance of the node. Lin proposed that the <table> tag is widely used to make the structure of a web page and it is using to extract blocks from a

web page [18]. The web pages basically created by HTML tags. The tags must be providing structural pattern of web pages. Such as <table>,
, and so on… the tag defines a division or a section in an HTML document and it is often used to group block-element. Mr. Patil used entropy based method the calculation of entropy value of a content block. It is given by summation of its features entropies. i.e.

$$H(CBi) = \sum_{j=1}^{k} H(Fj) \qquad (1)$$

Where Fj is a feature of CBi with k features. The equation can be normalized as content blocks contain different numbers of features

$$H(CBi) = \frac{\sum_{j=1}^{k} H(Fj)}{k} \qquad (2)$$

The entropy of a content block H(CB) , is the average of all entropy values in that block. Using this H(CB) a content block is identified as informative or redundant. If the H(CB) is higher than a threshold or close to 1 then content block is redundant as most of the block's features appear in every page. If H(CB) is less than a threshold then the content block is informative as features of the page are distinguishable. Yi, Liu and Li [4] consider non-content blocks as noise in the web page. They use a tree structure called Style Tree, to capture common presentation styles and actual contents of the pages in the given web site. A Style Tree can be built for the site by sampling the pages of the site. This tree is referred as Site Style Tree (SST).

Two data mining tasks approach web page clustering and classification the proposed technique was accessed. Experimental outcome illustrated that the mining results were enhanced considerably by the noise eliminating technique. Two algorithms are used content extractor and feature extractor algorithm. Feature Extractor is based on this characterization and uses heuristics based on the occurrence of certain features to identify content blocks. Content Extractor identifies non-content blocks based on the appearance of the same block in multiple web pages.

The Content Extractor algorithm eliminates blocks depending upon the inverse block-document frequency, IBDF, of a block. The inverse block-document frequency, IBDF, is inversely proportional to the number of documents in which a block occurs or has a similar block. Blocks that are similar to blocks occurring in multiple pages in the same domain, e.g. blocks that occur in multiple pages at cnn.com, are identified as redundant blocks.

Blocks that occur only in one page are identified as content-blocks. If S is a set of Web pages of the same class, i.e., obtained from the same source. Then

$$S = \{P1, P2, P3 \ldots PM\} \qquad (3)$$

Let us assume IBDF represents the IBDF of a block Bi in a set of pages S. Typically, the set S consists of similar pages from the same source. $IBDF^i$ Is inversely proportional to the number of web pages the block Bi occurs in. Then

$$IBDF^i \equiv f\left(\frac{1}{|S^i|+1}\right) \qquad (4)$$

Where,

$$S^i = \cup \{P_l : sim(B_i, B_k) < \in, \forall B_k \in P_l, \forall P \in S\}$$

f denotes a function, usually linear or log function. The function $sim\ (B_i : B_k)$ is a similarity measure of the two blocks. An expert provides the threshold∈. Given two blocks, the similarity measure, sim, returns the cosine between their block feature vectors. Examples of features are: the number of terms, the number of images, the number of java-scripts, etc. However, for text blocks, simply taking the number of terms in

the block may result in falsely identifying two blocks as similar. Therefore, we augment the features by adding a binary feature for each term in the corpus of documents.
If a feature occurs in a block, the entry in the corresponding feature vector is a one, otherwise it is zero. We used a threshold value of ∈ = 0:9. That is, if the similarity measure is greater than the threshold value, then the two blocks are accepted as identical.
The Feature Extractor algorithm is invoked to identify blocks with a set of desired features. Within the set of chosen blocks we sort the blocks again according to their probability values, and chose the winner block. Both content extractor and feature extractor produce excellent position and recall values and do not use any manual input and requires no complex machine learning process. They significantly outperform entropy based blocking algorithm proposed by Lin and Ho.

PROPOSED SYSTEM

The process of extracting informative blocks from a web page, a new proposed technique is based on the idea of web page segmentation and it is a semiautomatic proposed system. The input to the system is set of web pages and the output is set of informative content blocks within a web page. Following fig. is a sample images from BBC News with main contents, advertisements, navigation links etc.

Fig 1 A part of a web page with main contents and local noises

So the proposed technique is using to eliminate the non-informative content blocks and only shows informative content blocks in the output. The steps of the proposed system are described in fig 2. This research work proposes an approach for eliminating noise from web pages for the purpose of improving the accuracy and efficiency of web content mining. The main objective of removing noise from a web page is to improve the performance of the search engine.

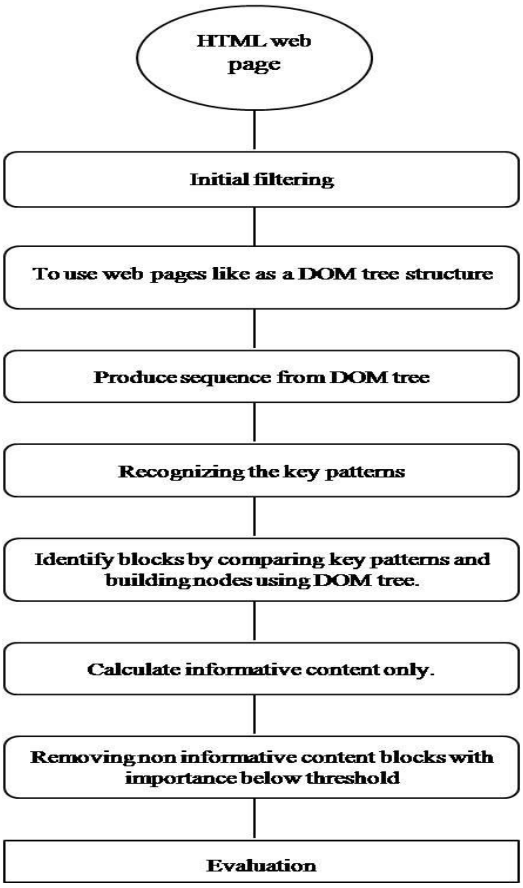

Fig 2 Proposed technique approaches for extracting the important blocks

A. INITIAL FILTERING

Noise elimination can be implemented as a filtering step for web content mining and especially for web page classification. Our objective is to find how to identify noisy blocks or irrelevant blocks from an input record, the web page to be processed, with a reduced complexity and increased efficiency. It is an important step for identify the informative content blocks in a page that leads to efficient information extraction. Processing step used to differentiate meaningful tags and less meaningful tags in Html source of the page. Each tags within a web page separated as basic tags, formatting tags, tags used for forms, frames, and images, audio, video, links, lists, tables, style, Meta info and programming based on their function. Less meaningful tags such as <a>, <b>, <script>, and <style> in the HTML source of the pages are removed.

B. TO USE WEB PAGES LIKE AS A DOM TREE STRUCTURE

After preprocessing, DOM tree structure nodes are either HTML tags or content consisting of texts and images. So in this process represent a web page as a DOM tree structure model. The Document Object Model (DOM) is a programming API for HTML and XML documents. It describes the logical structure of documents and the way a document is accessed and manipulated.

C. PRODUCE SEQUENCE FROM DOM TREE

This process uses only one-depth child nodes, ignoring other "deep" descendant nodes [2]. It has the advantage of reducing computational costs while still preserving some features of the DOM tree. For example, consider the sequence obtained after considering 1-depth child nodes and ignores deep descends for a sample web page is "h3 p h3 li h3 h3 p h3 h3 h3 h3 h3 h3 div" [2].

D. RECOGNIZING THE KEY PATTERNS

Mostly key patterns are repeating patterns in this section. These key patterns are longest and most frequent. In this section a repetition is defined as a subsequence of length m (>1) occurring twice or more

in a sequence of length n. According to this definition, repetitions maximum length for a sequence of length n is n/2, satisfying the formula of $1<m\leq n/2$. For Example fig. 3 repetition of length 2 and 3 produced for the given input sequence. In this section the only repeating pattern is AB as it occurs twice in the sequence [2].

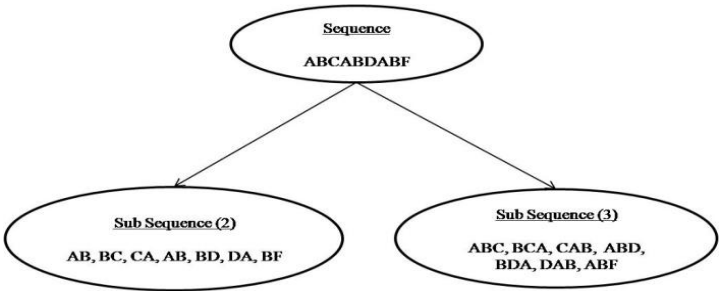

Fig 3 Repeating patterns from a sequence

When all the elements in a sequence are same, a repeating pattern must not be overlapped with other patterns in a sequence. The following sequence is obtained by considering 1 depth child nodes: "h3 p h3 li h3 h3 p h3 h3 h3 h3 h3 h3 div". The repeating patterns obtained here are [h3, p], [h3, h3], [h3, h3, p] and [h3, h3, h3]. Among these key patterns will be [h3, h3, p] and [h3, h3, h3], since [h3, p] and [h3, h3] are properly contained in [h3, h3, p] and [h3, h3, h3].

E. IDENTIFY BLOCKS BY COMPARING KEY PATTERNS AND BUILDING NODES USING DOM TREE STRUCTURE.

Consider a key pattern [h3, h3, p] and the sequence "h3 p h3 li h3 h3 p h3 h3 h3 h3 h3 h3 div". Here, the first match between the key patterns and the sequence occurs at position 1 and the second match occurs at position 7. Therefore, separate the two subsequences h3 p h3 li h3 h3 and p h3 h3 h3 h3 h3 h3 div. For each subsequence, a parent node is added with the parent node as the root and the element of the subsequence as its child node. Finally, from two key patterns, segment with two parent nodes and one child node are obtained as shown in fig. 4. In situation where all elements of a key pattern are the same, all subsequences matched with the key pattern are grouped within a dom tree structure.

For instance, the parent node separate with p1 and p2, and child node is separating with c1 and it is generated as result of matching between the key patterns [h3 h3 h3] and the sequence "h3 h3 h3 h3 h3 h3".

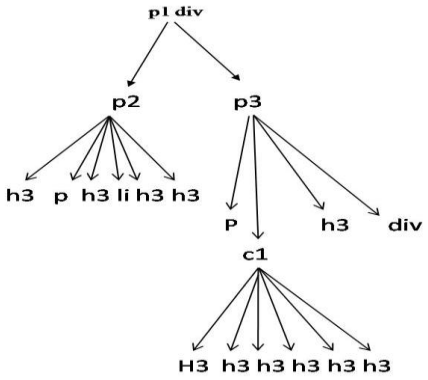

Fig 4 Generating nodes in tree structure

F. CALCULATE INFORMATIVE CONTENT.

During filtering approach all the less meaningful and unimportant tags are removed, remaining tags are identifying as an important tags. Important blocks are calculated by counting the number of important tags present in a block.

G. REMOVING NON INFORMATIVE CONTENT BLOCKS WITH IMPORTANCE BELOW THRESHOLD

The importance blocks are available in all blocks. Importance block < threshold (t) will be considered as non informative content blocks and it will be eliminated.

Important blocks selected in threshold level. The remaining blocks are eliminated from the web page as noisy blocks [1].

H. EVALUATION

The proposed system measuring the f score value. F score measures the performance of a system and it measured by finding precision and recalls value. "P" denotes the precision value that measures the ratio of correctly segmented blocks over the blocks segmented blocks over the blocks segmented by the proposed technique, and "R" denotes the recall value that measures the ratio of correctly segmented blocks over the ideal blocks that are manually obtained by humans. The F score measures and it reflects the average effect of both precision and recall [9].

$$P = \frac{\text{correctly segmented blocks}}{\text{blocks segmented by the proposed system}} \quad (5)$$

$$R = \frac{\text{correctly segmented blocks}}{\text{ideal blocks that are manually obtained by humans}} \quad (6)$$

$$\text{F score} = \frac{2*P*R}{P+R} \quad (7)$$

EXPERIMENTAL RESULTS

Finally the proposed system was tested in this section. The proposed technique is run against these pages and the results are assessed by experts. Following table 1 shows the precision, recall and F score. Fig 5 and Fig 6 shows precision and recall values graph and F score graph.

TABLE I

Precision, Recall and F Score

Page id	Precision	Recall	F score
1	0.66666	1	0.799995
2	1	0.7888	0.881932
3	0.98722	0.66666	0.795874
4	1	0.76888	0.869341
5	0.86745	0.5	0.634356
6	0.75	1	0.857143
7	0.7774	0.7764	0.7769
8	0.74433	1	0.853428
9	0.33333	0.777777	0.466663
10	0.666667	0.87676	0.757414
11	0.33333	0.888	0.484713
12	0.666666	0.777	0.717617
13	0.98	1	0.989899
14	1	0.7786	0.87552
15	0.89888	0.88888	0.893852
16	0.7869	0.8988	0.839136
17	0.8788	0.777777	0.825208
18	1	1	1
19	0.8888	1	0.941127
20	0.8341	0.66666	0.741039
21	0.786888	0.6756	0.72701
22	0.76888	0.4888	0.597654

23	0.87676	0.9898	0.929857
24	0.99999	0.76766	0.868557
25	0.87866	0.88888	0.88374
Average	**0.814868**	**0.825669**	**0.800319**

Fig 5 Graph of precision and recall

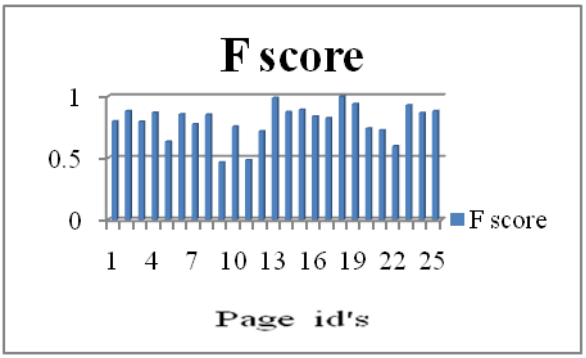

Fig 6 Graph of F Score

CONCLUSION

The proposed technique described in this paper for extraction of informative content blocks and elimination of non informative blocks by using Web page Segmentation. Here, a web page is divided into n blocks and the important blocks are calculated for each block. After removing the noise blocks, the remaining blocks considered as important blocks are extracted. Also the keywords from those blocks are extracted.

This process is based on the data in Table 1, it is observed that the average value of precision is 0.814868 and the average value of recall is 0.825669. The average value of F score is 0.800319. The proposed method provides better results in terms of Precision, Recall and F score [19].

REFERENCES

1. P. Sivakumar , R. M. S Parvathi , "An Efficient Approach of Noise Removal from Web Page for Effectual Web Content Mining", European Journal of Scientific Research ISSN 1450-216X Vol.50 No.3 (2011), pp.340-351 © Euro Journals Publishing, Inc. 2011
2. Jinbeom Kang, Jaeyoung Yang, Nonmember and Joongmin Choi, Member, IEEE "Repetition-based Web Page Segmentation by Detecting Tag Patterns for Small-Screen Devices", IEEE Transactions on Consumer Electronics, Vol. 56, No. 2, May 2010
3. S. H. Lin and J. M. Ho , "Discovering Informative Content Blocks from Web Documents", Proc. Eighth ACM SIGKDD Int'l conf. Knowledge Discovery and Data Mining , pp. 588-593, 2002.
4. Lan Yi, Bing Liu, Xiaoli Li, "Eliminating Noisy Information in Web Pages for Data Mining", SIGKDD .03, August 24-27, 2003, Washington, DC, USA.

5. Sandip Debnath, Prasenjit Mitra, C. Lee Giles, "Automatic Extraction of Informative Blocks from Webpages", SAC'05 March 2005, Santa Fe, New Mexico, USA
6. Lan Yi, Bing Liu, "Web Page Cleaning for Web Mining through Feature Weighting" SAC' 05 March 13-17, 2005, New Mexico, USA
7. Manisha Marathe, Dr. S.H.Patil, G.V.Garje,M.S.Bewoor, "Extracting Content Blocks from Web Pages", REVIEW PAPER International Journal of Recent Trends in Engineering, Vol 2, No. 4, November 2009
8. A. Arasu and H. Garcia-Molina, "Extracting structured data from web page," Proc. ACM SIGMOD Intl. Conf. on Management of Data, pp. 337–348, 2003.
9. Shine N. Das, Pramod K. Vijayaraghavan, Midhun Mathew, "Eliminating Noisy Information in Web Pages using featured DOM tree," International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868, Foundation of Computer Science FCS, New York, USA Volume 2– No.2, May 2012 – www.ijais.org
10. L. Yi, B. Liu, and X. Li, "Eliminating noisy information in web pages for data mining," Proc. 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, pp. 296-305, 2003.
11. D. Chakrabarti, R. Kumar, and K. Punera, "Page-level template detection via isotonic smoothing," Proc. 16th Intl. Conf. on World Wide Web, pp. 61-70, 2007.
12. Y. Chen, W.-Y. Ma, and H.-J. Zhang, "Detecting web page structure for adaptive viewing on small form factor devices," Proc. 12th Intl. Conf. on World Wide Web, pp. 225–233, 2003.
13. Y. Chen, X. Xie, W. Ma, and H. Zhang, "Adapting web pages for small screen devices," IEEE Internet Computing, vol. 9, no. 1, pp. 40-56, 2005.
14. Y. Yang and H. Zhang, "HTML page analysis based on visual cues," Proc. 16th Intl. Conf. on Document Analysis and Recognition, p. 859, 2001.
15. G. Hattori, K. Hoashi, K. Matsumoto, and F. Sugaya, "Robust web page segmentation for mobile terminal using content distances and page layout information," Proc. 16th Intl. Conf. on World Wide Web, pp. 361–370, 2007.
16. C. Choi, J. Kang, and J. Choi, "Extraction of user-defined data blocks using the regularity of dynamic web pages," Lecture Notes in Computer Science, vol. 4681, pp. 123-133, 2007.
17. S. Lin and J. Ho, "Discovering informative content blocks from Web documents," Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, pp. 588-593, 2002.
18. A. K. Tripathy and A. K. Singh, "An Efficient Method of Eliminating Noisy Information in Web Pages for Data Mining", In Proceedings of the Fourth International Conference on Computer and Information Technology (CIT'04), pp. 978 – 985, September 14-16, Wuhan, China, 2004.
19. Stevina Dias, "Identifying Informative Web Content Blocks using Web Page Segmentation", International Journal of Applied Information Systems (IJAIS)–ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 7– No. 1, April 2014.