International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 7– No. 1, April 2014 – www.ijais.org

Identifying Informative Web Content Blocks using Web Page Segmentation

Stevina Dias
M. E Student
TSEC, Mumbai, India

Jayant Gadge
Associate Professor
TSEC Mumbai, India

ABSTRACT

Information Extraction has become an important task for discovering useful knowledge or information from the Web. A crawler system, which gathers the information from the Web, is one of the fundamental necessities of Information Extraction. A search engine uses a crawler to crawl and index web pages. Search engine takes into account only the informative content for indexing. In addition to informative content, web pages commonly have blocks that are not the main content blocks and are called the non-informative blocks or noise. Noise is generally illogical with the main content of the page and affects two major parameters of search engines: the precision of search and the size of index

In order to improve the performance of information retrieval, cleaning of Web pages becomes critical. The main objective of proposed technique is to eliminate the non-informative content blocks from a Web Page. In the proposed technique, the extraction of informative content blocks and elimination of non informative blocks is based on the idea of Web page Segmentation. Here, a web page is divided into n blocks and the block importance is calculated for each block. The blocks with importance >=threshold are considered as important blocks and the remaining blocks are eliminated as noisy blocks. The proposed approach saves significant space and time

Keywords

Search engine, information extraction, web content mining, web segmentation, repetition detection, Informative blocks, non-informative blocks, and noise

1. INTRODUCTION

Web contents such as multimedia data, structured i.e. XML documents, semi-structured i.e. HTML documents and unstructured data i.e. plain text [1] offer valuable information to the users and can therefore be termed as informative contents. However, useful information on the Web is often accompanied by contents such as navigation bars, banner advertisements copyright notices etc [3] which can be termed as non informative contents.

The non-informative blocks are functionally useful for human viewers and essential for the Web site owners, but they often slow down automated information gathering and Web data mining tasks such as Web page clustering, classification/categorization, information retrieval and information extraction [4]. Therefore these blocks are termed as the noisy blocks. Also, from the users' perspective only part of the information is useful for a particular application and the remaining information are noises.

For improving the performance of traditional information retrieval, it is vital to differentiate valuable information from noisy content that may mislead users' notice within a solitary web page. Also, users barely pay notice to the commercials or copyright when they surf a web page. Information contained in these noisy blocks can seriously hamper Web data mining task. Eliminating these noisy blocks is thus of great importance.

This paper presents a technique to extract informative content blocks. Section II gives an insight to the various techniques to extract informative blocks. Section III describes the proposed technique. Section IV present results and Section V presents conclusion and future work.

2. LITERATURE SURVEY

Search engine uses a crawler that parse and index web pages and their content. Parsing and indexing the vast amount of information on the web is a serious challenge. One of the main issues on the web is that the information contained on web sites is often mixed with non-informative content, which might mislead crawlers on the semantics of the page [4]. Eliminating the non-informative blocks will help in improving the process of information extraction. One approach to identify the informative blocks is webpage segmentation. This section presents a summary of the techniques used to identify informative blocks within a web page.

The template-based method focuses on measuring the structural similarities among the DOM trees of Web pages. The template-based method builds a template with extracting rules organized by regular expressions [8]-[10]. It collects Web pages from a target site and generates regular expression rules in order to extract content blocks through analyzing the common zone. A pitfall of this method is 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. It also needs qualitative and quantitative training examples to cover real-world situations. To solve this problem, Chakrabarti proposed a method using a classifier [11]. However, it still needs many Web pages to train a classifier and build a template, causing the segmentation to be restricted. To overcome these difficulties, an approach was proposed by using the visual information in a Web page.

The vision-based method utilizes visual clues in a Web page. Chen proposed a method that considers visual information such as height, length of node zone, and separation information such as <HR> tag [12] [13]. Yang proposed the VIPS (Vision-based Page Segmentation) algorithm by considering vision information and heuristic rules to identify blocks [14]. The algorithm uses one essential term: Degree of Coherence or simply DoC. DoC is a measure of visual coherence defined for each block. It can be represented by any number integer or real, but it should grow with visual

consistency of the block. In addition a parent can never have greater DoC than its children in block hierarchy tree. The 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.

The tag-based method predefines content tags that contain useful information and finds content blocks by measuring the distance between these tags [13], [15], [16], [17]. In particular, Lin assumed that the <TABLE> tag is widely used to make the structure of a Web page, and proposed a method primarily using the <TABLE> tag to extract blocks from a Web page [18]. However, this method cannot be applied to those Web pages without the <TABLE> tags. To solve this problem, Debnath and Peng considered not only the <TABLE> tags, but also the <TR>, <P>, < HR>, and <UL> tags []. Basically, the layout of many web pages follows a similar pattern in such a way that the main content is enclosed in one big or <td> element which is HTML tags [5]. Some researchers concentrate only the content inside the "div" tag. The tag defines a division or a section in an HTML document and it is often used to group block-elements.

Many web sites use a HTML tag <TABLE> to layout their pages. Lin and Ho [3] used this observation to develop Infodiscoverer system to extract content blocks. In this method first a coarse tree structure is obtained by parsing a HTML page using a <TABLE> tag. Each internal node shows a content block containing one or more content strings as its leaf nodes. After parsing a web page into content blocks features of each block are extracted. Here features mean the meaningful keywords. After extracting features entropy value of a feature is calculated according to the weight distribution of features appearing in a page cluster. Next step is 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 [7]. Yi, Liu and Li [4] proposed a tree structure, called Style Tree, to capture the common presentation styles and the actual contents of the pages in a given Web site [7]. By sampling the pages of the site, a Style Tree can be built for the site, which is called the Site Style Tree (SST).

The definition of noise is based on the following assumptions: (1) the more presentation styles that an element node has, the more important it is, and vice versa. (2) The more diverse that the actual contents of an element node are, the more important the element node is, and vice versa [4]. Both these importance values are used in evaluating the importance of an element node. The presentation importance aims at detecting noises with regular presentation styles while the content importance aims at identifying those main contents of the pages that may be presented in similar presentation styles. Hence, in the proposed method the importance of an element node is given by combining its presentation importance and content importance. The greater the combined importance of an element node is, the more likely it is the main content of the pages. By employing two data mining tasks namely Web page clustering and classification the proposed technique was assessed. Experimental outcome illustrated that the mining results were enhanced considerably by the noise elimination technique.

Content-Extractor and Feature Extractor Algorithm The algorithms identifies primary content blocks by i) looking for blocks that do not occur a large number of times across web pages and ii) looking for blocks with desired features respectively. They identify primary content blocks with high precision and recall, reduce storage requirements for search engines, and result in smaller indexes. Performance evaluation shows that content extractor significantly outperforms the entropy based algorithm proposed by Lin and Ho in terms of accuracy and run-time

To design content-extractor algorithm Debnath et al [5] used the same basic concept used by Lin [2], that a <TABLE> tag is used to design maximum web pages. Unlike Lin they make use of some other html tags also while designing the algorithm. Similar blocks across different web pages obtained from different web sites can also be identified using this algorithm. In a table occurring in a web page, each cell is considered as a block. Where tables are not available, blocks can be identified by partitioning a web page into sections that are coherent. Many times news articles written by global news agencies appear in many news papers. User wants only one of these several copies of articles. These copies of articles differ only in their non-content blocks, so by separating non-content blocks from content blocks these same copies can be identified. As only unique articles are returned this will improve search results.

FeatureExtractor uses heuristics based upon the occurrence of certain features to identify content blocks. For example, FeatureExtractor invoked with the features text, image, links, identifies text blocks, image blocks or navigational blocks. Then it partitions the set of blocks into two partitions using clustering algorithm and selects the blocks that have desired features. Both FeatureExtractor and ContentExtractor produce excellent precision 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 [3] [5].

3. PROPOSED SYSTEM

In order to extract informative blocks from a web page, a new technique is proposed. The proposed technique is based on the idea of web page segmentation. The steps of the proposed system depicted in Fig. 1 and explained in detail in this section.

The Input to the system is Set of Web pages and the Output is Set of informative blocks within a web page.

3.1 Preprocessing

Pre-processing is an important step for recognizing informative content blocks in a page that leads to efficient information extraction. As a part of preprocessing, less meaningful tags in the HTML source of the page are removed. Tags within a web page can be as categorized as basic tags, formatting tags, tags used for forms, frames, and images, audio, video, links, lists, tables, style / section, Meta info and programming based on their function. These tags can be classified as meaningful and less meaningful tags. During preprocessing, less meaningful tags such as <a>, <b>, <script>, , and "#comment" in the HTML source of the page are removed.

3.2 Represent web page as a DOM tree

After the preprocessing step, a Web page is represented as a DOM tree structure whose nodes are either HTML tags or contents consisting of texts and images. The Document Object Model (DOM) is a programming API for HTML and XML documents. It defines the logical structure of documents and the way a document is accessed and manipulated.

3.3 Generate sequence from the DOM tree

A sequence is generated from the DOM tree of a Web page using only one-depth child nodes, ignoring other "deep" descendant nodes [2]. This approach of considering only one-depth child nodes has the advantage of reducing computational costs while still preserving some hierarchical features of the DOM tree. Hence, without considering descendant nodes with more than one depth from the root, the blocks generated by considering repetitions will remain consistent [2]. For example, consider the sequence obtained after considering 1-depth child notes and ignoring deep descends for a sample web page is "h2 p p ul p p h2 p p p p p p p div"[2]

3.4 Identify key patterns

A key pattern is a repetitive pattern is longest and most frequent. 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, the maximum length of a repetition for a sequence of length n is n/2, satisfying the formula of $1<m\leq n/2$ [2].

For example as shown in Fig 2 repetition of length 2 and 3 can be generated for the given input sequence. Here the only repetition is AB as it occurs twice in the sequence [2].

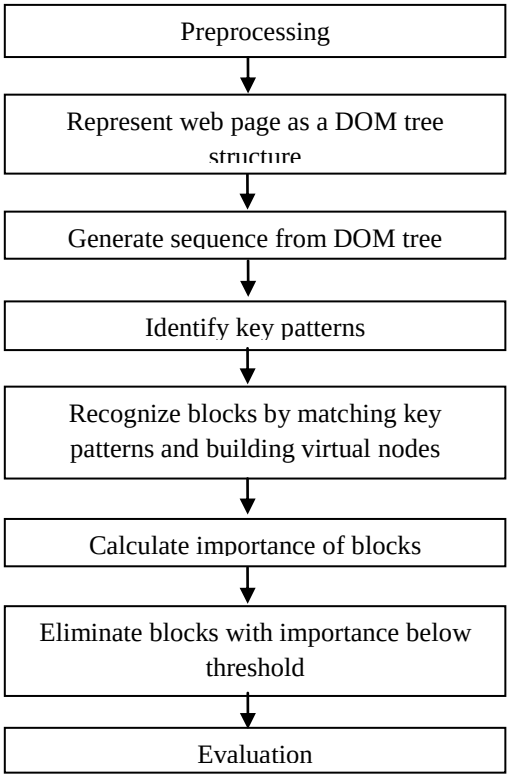

Fig 1: Proposed approach for extracting the important blocks

Sequence	Subsequence (2)	Subsequence (2)
A B C A B D	A B B C C A A B B D	A B C B C A C A B A B D

Fig 2: Repetitions from a sequence

When all the elements in a sequence are same, a repetition must not be overlapped with other patterns in a sequence. The number of repetitions of aa should be 2 as shown in case 2 [2] of Fig 3.

Fig 3: Repetitions from a sequence when all elements are same

Consider that the following sequence is obtained by considering 1 depth child nodes: "h2 p p ul p p h2 p p p p p p p div" [2]. The repetitions obtained here are [h2, p], [p, p], [h2, p, p] and [p, p, p]. Among these, key patterns will be [h2, p, p]

and [p, p, p], since [h2, p] and [p, p] are properly contained in [h2, p, p] and [p, p, p] [2].

3.5 Recognize blocks by matching key patterns and building virtual nodes

Consider a key pattern [h2, p, p] and the sequence "h2 p p ul p p h2 p p p p p p div". Here, the first match between the key pattern and the sequence occurs at position 1 and the second match occurs at position 7. Therefore, separate the two subsequences "h2 p p ul p p" and "h2 p p p p p p div." For each subsequence, a virtual node is added with the virtual node as the root and the elements of the subsequence as its children. Finally, from two key patterns, three virtual nodes 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 with a virtual node. For instance, the virtual node v3 is generated as a result of matching between the key pattern [p p p] and the sequence "p p p p p p."[2]

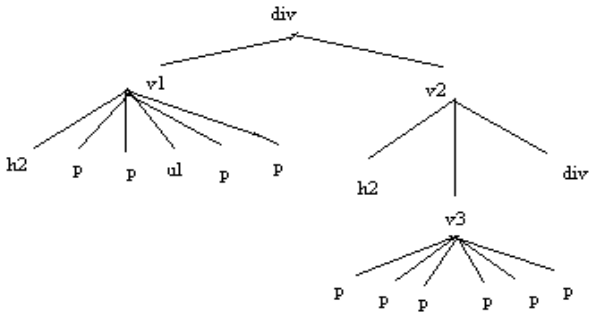

Fig: 4 Generating virtual nodes.

3.6 Calculate Block Importance

Since all the less meaningful and unimportant tags are removed during preprocessing, we assume that all remaining tags are important. Block importance is calculated by counting the number of important tags present in a block.

3.7 Eliminate noisy blocks i. e blocks with importance below threshold

All the blocks that have block importance < threshold (t) will be considered as noisy blocks and will be eliminated. The important blocks are selected only if the block importance is above the predefined threshold level. Otherwise; the remaining blocks are eliminated from the web page as noise blocks [1].

3.8 Evaluation

To ascertain the validity of the proposed measure, the proposed system is analyzed in terms of standard benchmark measures. "P" denotes the precision value that measures the ratio of correctly 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 the performance of a system and it reflects the average effect of both precision and recall [9].

$$P = \frac{\text{correctly segmented blocks}}{\text{blocks segmented by the proposed system}} \quad (3)$$

$$R = \frac{\text{correctly segmented blocks}}{\text{ideal blocks that are manually obtained by humans.}} \quad (4)$$

$$\text{F score} = \frac{2 * P * R}{P + R} \quad (5)$$

4. EXPERIMENTAL RESULTS

In order to test the proposed technique data set from cornell university is taken. The proposed technique is run against these pages and the results are assessed by experts. Table 1 shows the precision, recall and F score values obtained. Fig 5 shows the precision recall graph. Fig 6 shows the F score graph.

Table 1: Precision & Recall

Page id	Precision	Recall	F Score
1.	0.666667	1	0.8
2.	1	1	1
3.	0.75	0.33333333	0.461538
4.	1	1	1
5.	0.666667	0.5	0.571429
6.	0.75	1	0.857143
7.	0.666667	0.5	0.571429
8.	0.666667	1	0.8
9.	0.333333	0.5	0.4
10.	0.666667	0.66666667	0.666667
11.	0.666667	0.66666667	0.666667
12.	0.75	1	0.857143
13.	1	1	1
14.	1	0.4	0.571429
15.	0.5	1	0.666667
16.	0.5	1	0.666667
17.	0.666667	0.66666667	0.666667
18.	0.666667	1	0.8
19.	1	1	1
Average	**0.732456**	**0.80175439**	**0.73807596**

Fig. 5 Graph of Precision and Recall

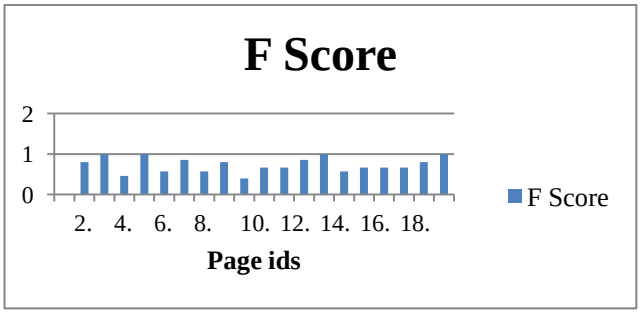

Fig. 6 Graph of F Score

5. CONCLUSION

The technique proposed in this paper for extraction of informative content blocks and elimination of non informative blocks is based on the idea of Web page Segmentation. Here, a web page is divided into n blocks and the block importance is calculated for each block. Using a predefined threshold value, the blocks with importance >=t are considered as important blocks and the remaining blocks are eliminated as noisy blocks.

Based on the data in Table 1, it is seen that the precision value varies from 0.5-1 and recall value varies from 0.3-1. It is also observed that the average value of precision is 0.732456 and the average value of recall is 0.80175439. The calculated F score varies from 0.4-1 and the average value of F score is 0.73807596. The proposed method provides better results in terms of Precision, Recall and F score.

6. 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 © EuroJournals 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.