Journal of Computer Science 7 (12): 1888-1893, 2011
ISSN 1549-3636

Sketching-Din Elimination of Web Page

[1]P. Sivakumar and [2]R.M.S. Parvathi
[1]Department of Computer Science and Engineering,
KSR College of Engineering,Namakkal, Tamilnadu, India
[2]Department of Computer Science and Engineering
Sengunthar College of Engineering,Namakkal,Tamilnadu,India

Abstract: Problem statement: The web content mining used to access lot of web pages, mining of web contents aims to extort positive information or awareness. **Approach:** There are several type of Web contents which can suggest valuable information to users are accessible in the Web, for instance graphical data, Extensible Markup Language documents, Hyper Text Markup Language documents and simple text. Here, only element of the information is useful for a testing purpose and the remaining information are noises. **Results:** In this research study, we propose an approach for removing the noises from a given web page which will get better the presentation of web content mining. At first, the web page information is divided into various blocks. **Conclusion:** From which, the duplicate blocks are removed using sketching. The performance of the proposed approach and results ensure the effectiveness of the proposed approach in classify the main blocks.

Key words: Web mining, web content mining, web cleaning, duplicate blocks, web page information, graphical data, world wide web, Web Structural Mining (WSM), Web Usage Mining (WUM)

INTRODUCTION

The World Wide Web is quickly promising as a significant standard for transacting trade as well as for the distribution of information allied to a large collection of topics for example industry, administration, Games. According to mainly prediction, the mass of person information will be accessible on the Web in ten years. These vast amounts of data raise a grand challenge, namely, how to turn the Web into a more useful information utility. Web content mining face huge problems due to the duplicate and near duplicate web pages. These pages either increase the index storage space or slow down or increase the serving costs thereby irritating the users. Thus the algorithms for detecting such pages are inevitable for effective web content mining.

Analysis and discovery of useful information from World Wide Web poses a phenomenal challenge to the researchers in this area. Such a phenomena of retrieving valuable information by adopting data mining techniques is called Web mining. Web mining is classified into following five sub tasks: (1) Resource finding, (2) Information selection and pre-processing, (3) Generalization, (4) Analysis and (5) Visualization (6).

Web mining is separated into three category: Web Content Mining (WCM), Web Usage Mining (WUM) and Web Structural Mining (WSM). Web content mining is the method of identify user definite data from text, image, audio or video data already available on the web. This process is alternatively called as web text mining, since text content is the most widely researched subject on the World Wide Web.

This is so due to the following characteristics of the Web: (1) the amount of data/information on the Web is huge and still growing rapidly. (AlMurtadha *et al*., 2011; Al Shalabi, 2009)Web data is also easily accessible, (2) the coverage of Web information is wide and diverse. One can find information about almost anything on the Web, (3) Information on the Web is heterogeneous, (4) Much of the Web information is semi-structured due to the nested structure of HTML code and the need of Web page designers to present information in a simple and regular fashion to facilitate human viewing and browsing, (5) Much of the Web information is linked. There are links among pages within a site and across different sites. These links serve as an information organization tool and also as indications of trust/authority in the linked pages and sites, (6) much of the Web information is redundant. The same piece of information or its variations may appear

Corresponding author: P. Sivakumar, Department of Computer Science and Engineering, KSR College of Engineering, Namakkal, Tamilnadu, India

in many pages or sites. This property has been explored in many Web data mining tasks, (7) The Web is noisy. A Web page typically contains a mixture of many kinds of information, e.g., main content, advertisements, navigation panels, copyright notices.

For a particular application only part of the information is useful and the rest are noises (Nasri *et al*., 2008). Web content mining is related but different from data mining. It is related to data mining because many data mining techniques (e.g., clustering, classification, association rules) can be applied in Web content mining. However, it is also quite different from data mining because Web data are mainly semi-structured and/or unstructured, while data mining deals primarily with structured data. Web content mining considers different kinds of data such as: images, audio, video and texts (e.g., web documents and free texts). For web documents, the mining methods are mainly focused on information extraction and integration (i.e., gathering explicit information from different web sites for its access) (6). In the web, there are different kinds of Web content which can provide useful information to users, for example multimedia data, structured (i.e., XML documents), semi-structured (i.e., HTML documents) and unstructured data (i.e., plain text).

Information in a web page is not uniformly significant. For example, consider the web page in Fig. 1, the caption in a news web site is much more attractive to users than the navigation bar. And users only just pay attention to the advertisement or the copyright when they browse a web page (Song *et al*., 2004).

Therefore, dissimilar information inside a web page has dissimilar importance weight according to its location, occupied area, content. Thus, it is off to assign importance to a region in a web page, we first need to segment a web page into a set of blocks.

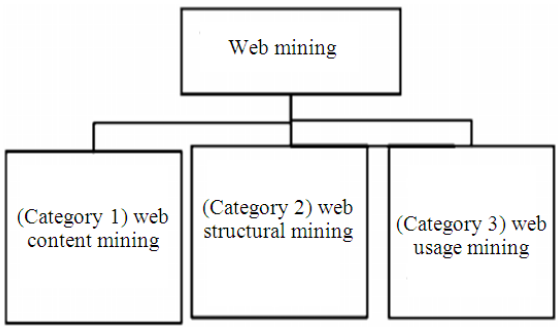

Fig. 1: Web mining categories

The programs that navigate the web graph and retrieve pages to construct a confined repository of the segment of the web that they visit are known as web crawlers. Earlier, these programs were known by diverse names such as wanderers, robots, spiders, fish and worms, words in accordance with the web imagery Web crawling is useful for a variety of purposes. Describing and organizing the vast amount of content available in the web is essential for realizing its full potential as an information resource. Powerful search engines have been developed to aid in locating unfamiliar documents by category, contents, or subject. The web is a huge repository of information and there is a need for categorizing web documents to facilitate the indexing, search and retrieval of pages. However these resources have been created by large teams of human editors and represent only one kind of classification task that, while widely useful, can never be suitable to all applications. Web page classification involves the classification of Web pages under some predefined categories that may be organized in a tree or other structures. Web clustering involves the grouping of Web pages based on the similarities among them.

Each resultant group should have similar Web pages while Web pages from different resultant groups should be dissimilar. In recent times, most of the available web documents are bound to contain noisy and irrelevant information in addition to the significant information. Moreover, a worrying number of documents available in WWW are exact or near exact duplicates of others. Web content mining includes the task of organizing and clustering the documents and providing search engines for accessing the different documents by keywords, categories, contents (Fu *et al*., 2007),Two main research focuses on web content mining have been paid more attention in the past few years. One is information extraction; the other is web page classification or clustering.On the dawn of the WWW, finding information was done mainly by scanning through lists of links collected and ordered by humans according to some criteria.

Noise eliminationl: The major purpose for removing noise from a Web Page is to get better the presentation of the search engine and differentiate important information from noisy content. In this research work are mainly absorbed to remove the noises: (1) Panels and frames, page headers and footers, advertisements and audio, video, duplicate contents and noise Contents according to block importance. The removal of these noises is done by block splitting operation, primary noises removing and the useful text contents are partition into blocks, Then Next step, using sketching, it's used to duplicate blocks are removed to take the different blocks. Using the remaining blocks with high importance is considered as important blocks and the keywords are extracted from those important blocks.

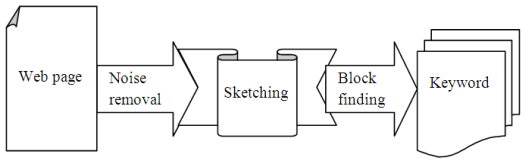

Fig. 2: Important block finding method

The Fig. 2 explains the in general procedure of noise removing web page for web mining. The different stages of our proposed work are as follows:

- Noise removing
- Performing sketching
- Importance block finding
- Keywords from the web page

MATERIALS AND METHODS

Block splitting: The web page contains slightest interested contents which are referred as main noises. Removing these noises will help in improving the mining of web. To allocate importance to an area in a web page, we first need to part a web page into a set of blocks. For this reason to clean a web page, a preprocessing step called block splitting operation is performed.

Fundamentally, the describe of many web pages follows a related model in such a way that the main content is enclosed in one big or <td> element which is HTML tags. In our research work, we are focusing 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. The block splitting procedure aim at cleaning the home noises by in view of only the major content of a web page with this in div tag. The major contents we take are divided into various blocks.

Block splitting algorithm:

Step 1: Select web page
Step 2: Identify local and primary web page
Step 3: Find out Local Noise
Step 4: If Tag EQL <DIV> or <TD> Then
Read content
Else
Exit
Step 5: If identified <DIV> or <TD> tag then
Stored content in block {"prathik", priya", "siva" }

$$B = \{B_1, B_2, B_3, \cdots, B_n\}, B \in W$$

Step 6: continue study up to end of web page
Step 7: Exit
Where, B = A set of blocks in the web page W
n = Number of blocks in a web page W

In Fig. 3 we have taken an example of a sample web page which consists of local noises such as images, multiple links and also the main content useful for mining. The dotted lines represented in the Fig. 4 are denoted as local noises. The useful main contents for web content mining are highlighted with dark lines. The main content has some sub-contents which are divided into blocks B_1 and B_2 using Block Splitting Operation.

Selecting distinct blocks using sketching algorithm: The sketching algorithm is used to find duplicate blocks, it's used to remove duplicate block from the web page. For finding the duplicate blocks, we have used fingerprint method proposed by Charikar, in this sketching algorithm is used to remove near duplicate content in the web page. (1) Mirror sites (2) FAQs, manuals, legal documents (3) Different versions of the same document (4) Plagiarism.

The process of generating a duplicate content is given as follows:

Step 1: U: space of all possible documents
Step 2: S $\subseteq$U: collection of documents
Step 3: Sim: U×U= (0,1): a similarity measure among documents
a) If p,q are very similar sim(p,q) is close to 1
a) If p,q are very unsimilar, sim(p,q) is close to 0
b) Usually: sim(p,q) = 1–d(p,q), where d(p,q) is a normalized distance between p and q.
Step 4: G: a graph on S:
a) p, q are connected by an edge iff sim (p, q) $\geq$t (t = threshold)
Step 5: Goal: find the connected components of G

The final result we obtain in this process is a set of distinct blocks which is represented as follows:

$$B_d = \{B_{d1}, B_{d2}, B_{d3}, \cdots, B_{dm}\},\ B_d \in B$$
$$B_{di} \in B_d ; \quad i = 1,2,3,\cdots,m$$

Where:
Bd = A set of distinct blocks
m = Number of distinct blocks

Block importance representation: Web page designer lean to arrange their satisfied in a logical way: giving importance to main effects and deemphasizing the insignificant part through suitable skin such as position, size, color, word and image. A block weight form is a purpose to map from features to meaning for each block and can be formalized as:

<block features> ® block importance

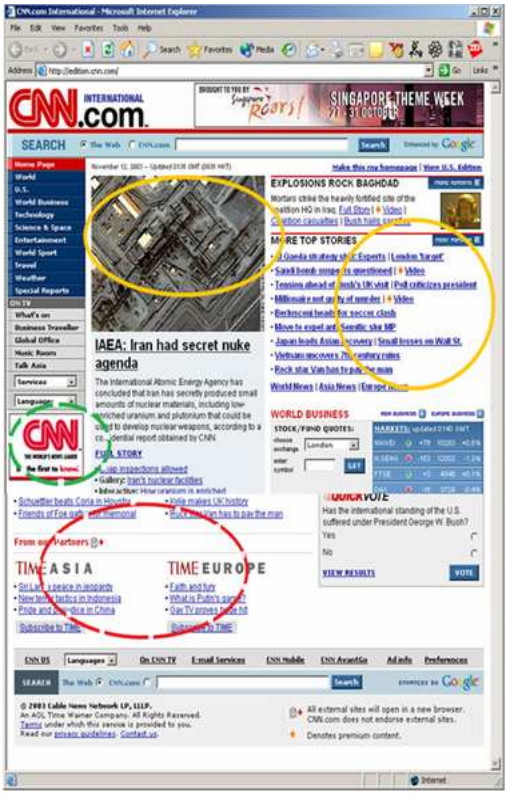

Fig. 3: A sample web page containing multiple regions with different importance

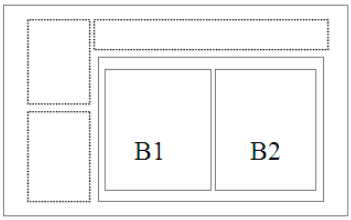

Fig. 4: Blocks with noises and main contents

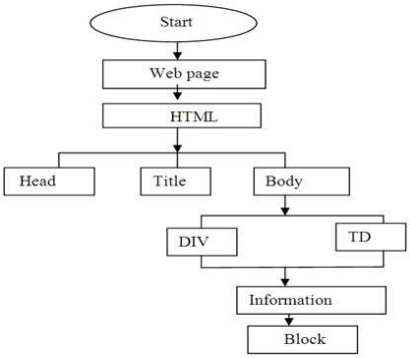

Fig. 5: Block splitting operation

Block features: We look up again the web page in Fig. 1. What type of concept is used to identify the important parts from unimportant parts? Web page designer is most important message is given into middle of the web page; put the navigation bar on the header or the left side and the copyright on the footer, the Fig. 1 is marked important block is solid circles and unimportant block is marked dashed circle. Thus, the importance of a block can be reflected by spatial features like position, size. On the other hand, the contents in a block are also useful to judge block importance. For example, the spatial features of both of the two solid circles in Fig. 1 are similar. But one contains a picture, a highlighted title and some words to describe a news headline and another contains pure hyperlinks pointing to other top stories. Based on the contents of the blocks, it is possible to differentiate their importance. Spatial features of a block are made up of four features:

{BlockCenterX, BlockCenterY,
BlockRectWidth,BlockRectHeight}
BlockCenterX and BlockCenterY are the coordinates of the center
point of the block and
BlockRectWidth, BlockRectHeight are the width and height of the block.

Such spatial features are called absolute spatial features since they directly use the absolute values of the four features. But using absolute values may make it hard to compare the features from different web pages. For example, a big block in a small page will always be taken as small block when comparing it with the blocks in a big page. So, by using the width and height of the whole page to normalize the absolute features, we transform them into relative spatial features, as given below:

{BlockCenterX/PageWidth, BlockCenterY/PageHeight,
BlockRectWidth/PageWidth,
BlockRectHeight/PageHeight}.

We found that size normalization brings up another problem. For some long pages with height times larger than the screen height (e.g., the page in Fig. 1 or pages longer than it), after normalization, some important blocks on the top part (i.e., blocks displayed in the first screen, such as the blocks in the solid circles in Fig. 1) may be transformed into blocks located at the top of the page with quite small height. In these cases, the spatial features of these important blocks are very similar to the spatial features of the unimportant blocks such as advertisements in short pages. The point here is that, for a long page, the content in the first screen is most important and we should avoid normalizing them with the height of the whole page. Width normalization does not have the same problem since few pages have widths bigger than the screen (Song *et al*., 2004).

In this Fig. 5 is representing operation of Block spiting concept, here first read the web pages and then convert HTML code, next find out head, title and body tag. After that divide div and TD tag. Finlay retrieves extract message form open div tag and close div tag to reach Block,

RESULTS AND DISUCSSION

Experimentation: The experimental results of the proposed approach for removing the noises from web pages are presented in this study. The proposed approach has been implemented in java (jdk 1.6) and the experimentation is performed on a 3.0 GHz Pentium PC machine with 2 GB main memory. For experimentation, we have taken many web pages which contain all the possible noises.

These pages are then applied to the proposed approach for removing the different noises. The removal of noise blocks and extracting of useful content blocks are explained in this sub-section. Finally, evaluation results ensure that our proposed approach acquired better results. The example results of four web pages are given in the following sub-section.

Evaluation results: In testing, we have taken the web page W_1 named "Advances in artificial neural" and this web page contained images which are noises that had to be removed first in order to separate the main contents from those noises. Hence, only the content inside the HTML tag "div" is considered as our main content. And these contents are divided into 4 various blocks using block splitting operation. Thus in this step we obtained a set of 4 blocks. In the second step, we performed sketching method for eliminating the duplicate blocks.

Evaluation result 1: Since, this Table 1 and 2 block spiting concept is given and also important block point is mentioning

Evaluation result 2:

Experimental setup:

Data: For our research we chose one web page with primary nosic and local nosic the total pages among the datasets are shown in Table 1.

Hereby given actual web page name and also that web page URL address it's used for easy identify people.

Block splitting operation: Here identified number of block with their imparted factor, in this method is very useful for content retrieval method; here by display all type of block in web page.

Table 1: Experiment dataset

Web page name	URL
Advances in artificial neural Systems-an open access journal	www.guidlines.html

Table 2: Block splitting operations

Block name	Block importance point
B1	0.63
B2	0.67
B3	0.60
B4	0.63
B5	0.80
B6	0.65
B7	0.60
B8	0.97
B9	0.67

Table 3: Duplicate block elimination

Block name	Block importance point
B1	0.63
B2	0.67
B3	0.60
B5	0.80
B6	0.65
B8	0.97

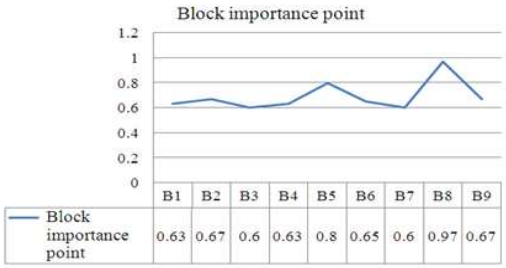

Fig. 6: Important block splitting

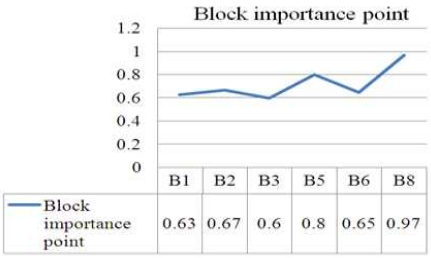

Fig. 7: Reduced duplicate block

CONCLUSION

To improve the quality of web content mining, we have proposed an effective approach which is used to eliminate the noises from the web pages (Table 3). The primary noises are considered as irrelevant data which can be removed by simple block splitting operation (Fig. 6). The duplicate blocks Fig. 7, which are now free from primary noises, can be removed by computing the fingerprint of each block using the Sketching algorithm. The Sketching algorithm is

preferred here since the efficiency of the algorithm is maintained when the number of entries is large. These parameters represent the importance of each block. The noise blocks represent the audio and video files of a webpage. They can be removed by using the threshold value. Apart from the noise blocks, we have considered the other blocks as important blocks and extracted the keywords from those blocks. From our proposed theory, it is easily understood that we have removed all the possible noises from the web pages under examination. Thus efficient web content mining is achievable.

REFERENCES

AlMurtadha, Y., N.B. Sulaiman, N. Mustapha and N.I. Udzir, 2011. IPACT: Improved web page recommendation system using profile aggregation based on clustering of transactions. Am. J. Applied Sci., 8: 277-283. DOI: 10.3844/ajassp.2011.277.283.

Al Shalabi, L.A., 2009. Improving accuracy and coverage of data mining systems that are built from noisy datasets: A new model. J. Comput. Sci., 5: 131-135. DOI: 10.3844/jcssp.2009.131.135

Fu, Y., D. Yang, S. Tang, T. Wang and J. Gao, 2007. Using XPath to discover informative content blocks of web pages. Proceedings of the 3rd International Conference on Semantics, Knowledge and Grid, Oct. 29-31, IEEE Xplore Press, Shan Xi, pp: 450-453. DOI: 10.1109/SKG.2007.106

Nasri, M., S. Shariati and M.A. Azgomi, 2008. Performance modeling of a distributed web crawler using stochastic activity networks. Communi. Comput. Inform. Sci., 9: 535-542, ISSN: 1865-0929.

Song, R., H. Liu, J.R. Wen and W.Y. Ma, 2004. Learning block importance models for web pages. Proceedings of the 13th international conference on World Wide Web, ACM New York, NY, USA., pp: 203-2011. DOI: 10.1145/988672.988700.