International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 2, February - 2014

Elimination of Noisy Information from Web Page using DOM and Ant Colony Optimization

Shaikh Sakina Banu
Department of Computer Engineering
Dwarkadas . J. Sanghvi College of Engineering
Mumbai, India

Hitesh Kumar Bhatia
Department of Masters of Computer Application
Sardar Patel Institute of Technology
Mumbai, India

***Abstract*— A webpage is a collection of different informational block which contains a lot of information including relevant information, and irrelevant information as well such as advertisements and navigation. In a given Web site, noisy blocks usually share some common contents and presentation styles, while the main content blocks of the pages are often diverse and different in their actual contents and/or presentation styles. We need to mine the relevant information from the web page to improve the performance of mining. For doing this DOM tree structure was used where information is selected, noisy information is marked and pruned .In this paper we propose a solution to eliminate noisy information from web page using DOM tree structure and ant colony optimization to improve the efficiency of mining. The different web pages are used to first construct DOM Tree. Our approach is based on finding noise in current web page and also web noise similarity by using Ant Colony Optimization (ACO) approach. We also apply neural network algorithm to categorize the stored various noise model by corresponding noise data in current web page.**

Keywords*— *Ant Colony Optimization;DOM(DocumenObject Modelling); noisy information; Neural Networks.

I. INTRODUCTION

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. The amount of information on the Internet shows exponential growth, and the size and number appear to be growing rapidly at a faster rate. Given the enormous volume of web pages in existence, it comes as no surprise that Internet users are increasingly using search engines and search services to find specific information. Data mining on the Web thus becomes an important technique for finding useful knowledge or information from the Web. However, useful information or knowledge on the Web is often accompanied by a large amount of noise such as banner advertisements, navigation bars, copyright notices, etc. Information present on the web page are useful for an individual but it hampers the process of information gathering due to present of various kinds of noises on that web page.

Web noise can be categorized into two types: global noise and local noise [1] [2]. Global noise covers almost the whole web page, for e.g. Web pages that are not updated for long time and has the older version of pages which means that it contains information which is of no use. Local noise includes all the information which has no significance with the web page for e.g. .advertisement, banners etc. Local noise resides within a web page.

DOM trees can be constructed easily as it remains highly editable and can easily be reconstructed back into a complete webpage. The DOM tree is hierarchically arranged and can be analyzed in the form of sections or as a whole, providing a wide range of flexibility. By parsing a webpage into a DOM tree, more control can be achieved while eliminating noise data. The first step of building block or model is to segment a web page to multiple regions or blocks. Several methods are available to segment a web page into blocks. In the DOM-based segmentation approach, an HTML document is represented as a DOM tree. Useful tags that represent a block or module in a page include P (for paragraph), TABLE (for table), UL (for list), H1~H6 (for heading), etc [5].

DOM in general provides a useful structure and better representation for a web page. But some tags are used not only for content organization, but also for layout presentation such as TABLE, P etc. In many cases, DOM tends to reveal presentation structure other than content structure, and is often not accurate. By parsing a webpage into a DOM tree, it has been found that one not only gets better results but has more control over the exact pieces of information that can be changed or modified while extracting content.

II. RELATED WORK

Elimination of noisy and irrelevant contents from web pages has many applications, including web page classification, clustering, web featuring, proper indexing of search engines, improving the quality of search results and text summarization. Thus cleaning web pages for web data extraction becomes mandatory for improving the performance of information retrieval. We are focusing to remove various noise patterns from web pages instead of extracting relevant content blocks from web pages.

Lan Yi et al. [4] proposed a compressed structure tree (CST) to capture the common structure and comparable blocks in a given set of web pages. It then evaluates the importance of each node in CST. Based on the tree and its node importance values, a weight is assigned to each word feature in its content block. The resulting weights were used in web mining, however it did not give accurate result.

T. Sun et al. Created a DOM tree on the visual blocks of a web page and for each block, an information block matching ratio is calculated. This ratio is matched with the threshold level to identify the level of relevancy. But the technique was mainly based on an assumption that the same site are

often made from a different page with an HTML template generation and their structure is very similar[5].
Majority of the techniques were based on the observation and assumption that web pages usually share some common layouts and presentation styles, which is not true in all cases especially when loading dynamic web pages. Also many of these techniques need a set of web pages even from a single web site, which is an extra burden while dealing with individual pages for web mining on the same website.
The HTML tag looks like given below:

```
<BODY bgcolor=GREEN>
<TABLE width=650 height=500 >
.
.
</TABLE>
<IMG src="grep.jpg" width=650>
<TABLE bgcolor=BLACK>
.
.
</TABLE>
</BODY>
```

The DOM Tree created for the above HTML structure is given in fig 1.

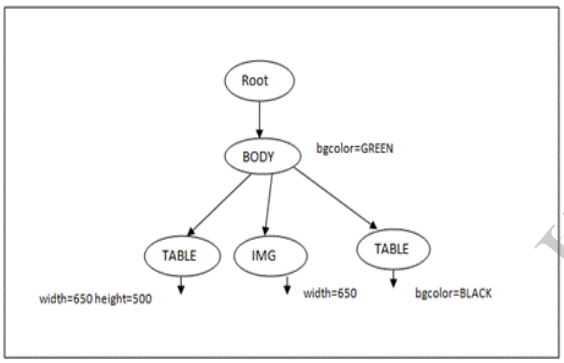

Fig 1: construction of Dom tree

The Dom tree is constructed in the above manner where each tag is considered as one node .there are various methods used for eliminating noise which are described below:

A. Featured DOM Tree

A three stage algorithm is proposed in which feature selection is done in the first phase, a featured DOM tree is created in the second phase and noise is marked and pruned in the third phase [3]. In the first phase which is featuring phase, standard web page pre- processing methods like html tag removal, tokenization, removal of stop words and stemming are applied on the input record and a feature set " F" of m tokens {x1,x2,….,xm} are retrieved.

In the second phase, modeling phase, the HTML document is modelled as a DOM tree. Each HTML page corresponds to a DOM tree where tags are internal nodes and the detailed texts, images or hyperlinks are the leaf nodes.DOM tree in general is sufficient for representing the layout or presentation style of an HTML page, however it is hard to study the content or semantics of the page to clean it. Therefore they introduced a new tree structure, called featured DOM tree, which is able to represent the presentation style as well as the feature sets of individual blocks of the web page. For creating a featured DOM tree, an optimal feature selection is done for individual leaf nodes of the DOM tree and feature weighting can also be applied here based on the leaf node tag. Similarity verification is done in the third phase and noisy blocks are marked, propagated and further eliminated. The weight percentage of each token in a feature set is calculated with respect to the total weight of the set and a new technique known as Minimum Weight Overlapping (MWO) is applied here for similarity verification.
Next step is to remove noisy blocks from DOM tree. For that purpose, a bottom up traversal is done on the tree in such a manner that a parent node is marked as a noisy one if all of its children are noisy otherwise the node is considered as relevant. So this marking can be propagated up the tree till we reach the parent node. Finally the marked portion of the DOM tree is pruned and remaining tree structured is mapped back into HTML page so that a cleaned web page can be obtained.

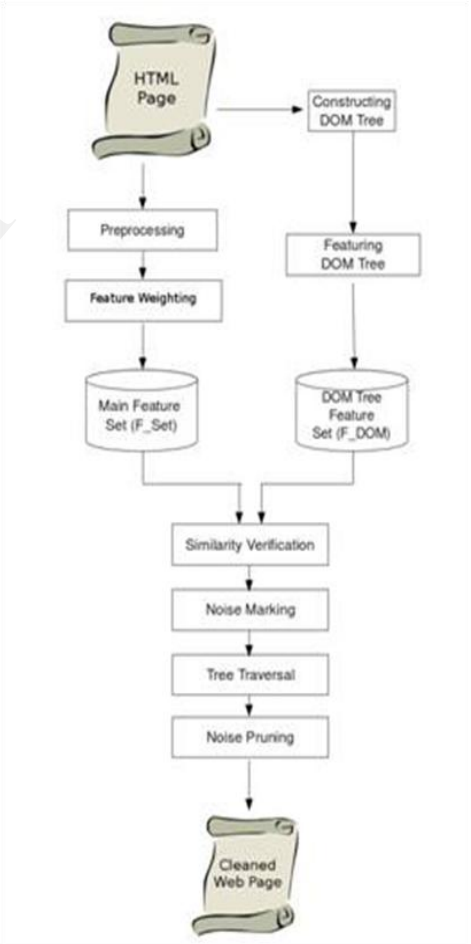

Fig 2: Overall flow diagram

B. Style Tree Structure

A new tree structure, called Style Tree, is proposed to capture the actual contents and the common layouts of the pages in a web site. A new content extraction method is thus proposed, which can discover web page content according to the number of punctuations and special markings and the ratio of non-hyperlink character number to character number that

hyperlinks contain. It can eliminate noisy information and extract main content block or relevant information from web pages effectively.

In this paper [5] they propose a novel idea for finding near duplicates of an input web page, from a huge repository. This approach explores the semantic structure, content and context, of a web page rather than the content only approach thus making the search more effective.A possible application of Neural Networks [4] is presented for three pattern classification combine with DOM structure to extract content information. The type of Neural Network used to implement the system is feed forward which uses the back propagation learning algorithm.

Different Web Sites have different layout and presentation style, therefore the depth of the tree of the Web page is varied according to their presentation style. The system must know the maximum level of DOM tree to choose the good choice of threshold level. Therefore, the system traverses the whole DOM tree to get the maximum depth of DOM thus making it less efficient.

C. *Case Base Reasoning*

CBR is a machine knowledge method that adopts an idle knowledge approach and contains no plain model of the difficulty area. Case-Based Reasoning (CBR) utilizes ancient times experiences as a key data source for future problem solving and is measured a new method in the improvement of Artificial Intelligence.

CBR uses periodic experiences stored in cases as a beginning for decisions and has been implemented in a wide range of fields. Each case is made up of a description of a past example or experience and its respective solution as it does consider previous history. The full set of past experiences encapsulated in individual cases is called the case base. The idea is to learn from experience or training. However, a crucial aspect of CBR lies in the term “similar”. When a new problem is presented, the case base is searched, similar past examples are found and these are used to solve the presented problem. Thus, we identify variety of noise patterns in many Web sites and these noise patterns in each site are represented by DOM (Document Object Model) tree and keep them into database as a case in our first task. DOM trees remain highly editable and can easily be reconstructed back into a complete webpage.

However there are various drawbacks using case based reasoning such as:

- Can take large storage space for all the cases.
- Can take large processing time to find similar cases in case-base.
- Cases may need to be created by hand.
- Adaptation may be difficult.
- If you require the best solution or the optimum solution - CBR may not be good.

III. PROPOSED ARCHITECTURE

Various Researchers have developed a number of approaches for retrieving and removing main content from Web pages. Most of them have listening carefully on identifying main content blocks in Web pages. Even though clean-up noisy data is an important task, reasonably small work has been complete in this playing field.

A. *Outline of Proposed Approach*

In this part, we initially demonstrate the two mechanisms that decide which area of present Web page contains noise or combination. Then, we plan another method on matching to determine how we process the three classes (noise, data and mixture) in ant colony optimization. Lastly, we remove the various noise patterns in current Web page and show extracted main content data.

1) *Ant Colony Optimization*

Ant colony optimization is a biologically inspired optimization method which gives efficient output through analysis. The basic idea is to use a large number of simple agents called ants where each ant performs a relatively simple task but combined together they are able to produce sophisticated and effective optimization.

Artificial ants have several characteristics similar to real ants, namely:

- Artificial ants have a probabilistic preference for paths with a larger amount of pheromone.
- Shorter paths tend to have larger rates of growth in their amount of pheromone.
- The ants use an indirect communication system based on the amount of pheromone deposited on each path.
- ACO algorithms are based on the following ideas:
- Each path followed by an ant is associated with a better solution for a given problem.
- When an ant follows a path, the amount of pheromone deposited on that path is proportional to the quality of the solution for the target problem.
- When an ant has to choose between two or more paths, the path(s) with a larger amount of pheromone have a greater probability of being chosen by the ant.

ACO algorithm has many advantages which overcomes the disadvantages of case based reasoning (CBR) such as:

- Can be used in dynamic applications
- Positive Feedback leads to rapid discovery of good solutions.
- Inherent parallelism.
- Accurate result of data.

2) *Artificial Neural Network*

An artificial neural network (ANN), often just called a "neural network", is a mathematical model or computational model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using some kind of computation. In most cases an ANN is an adaptive system that changes its structure based on input and output information that flows through the network during the learning phase.

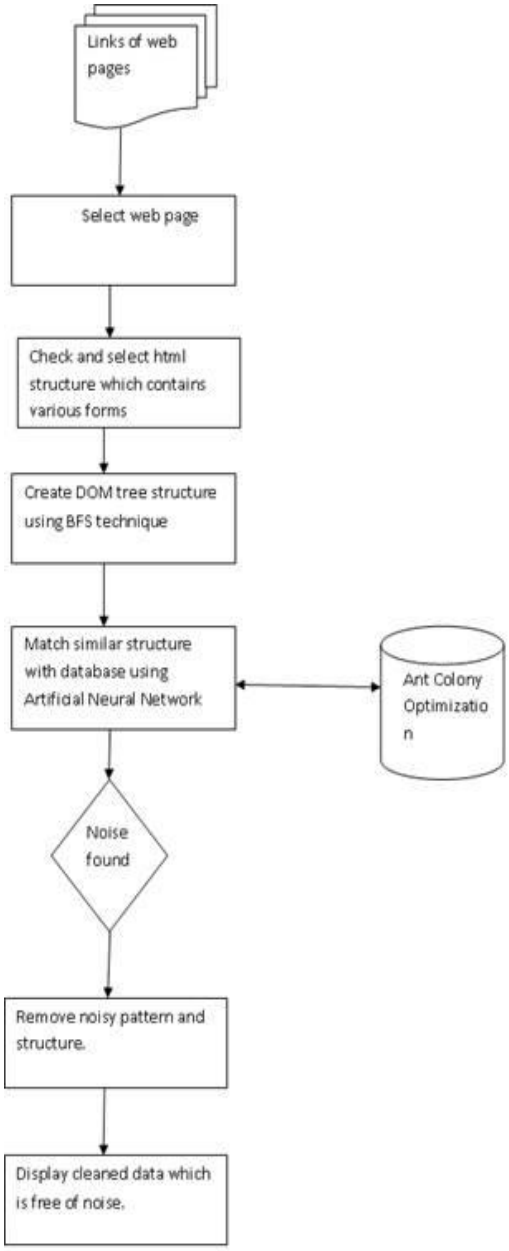

Figure 4: Outline of proposed approach

In our proposed work, we applied Artificial Neural Network model for pattern matching. The greatest strength of neural networks is their ability to learn by training. Artificial Neural networks are very good at pattern recognition and pattern-matching tasks. If the input is one it has never seen before, it produces an output similar to the one associated with the closest matching training input pattern. Neural network take input from the Dom tree and database where we have used ant colony optimization to get better result and then it matches for the pattern for noisy information once it gets the pattern it removes it or eliminates it.

IV. ALGORITHM FOR NOISE REMOVING

Input: Multiple web Documents

Method: Ant Colony Optimization Method

Output: Extraction of relevant web documents free of noise.

Step 1: Access multiple web page

Step 2: Read one by one page

Step 3: Check Web HTML tag

Step 4: Consider the document with various tags

Step 5: create DOM Tree structure using HTML parser.

Step 6: Train the dataset in database where all the information related to web pages is stored for efficient retrieval of pattern by using Ant Colony Optimization technique.

Step 7: Match the constructed DOM tree with the information in the database using Artificial Neural Network and retrieve similar kind of information or pattern which contains noisy data and eliminate it from each web page.

Step 8: Finally receive web page without noisy data

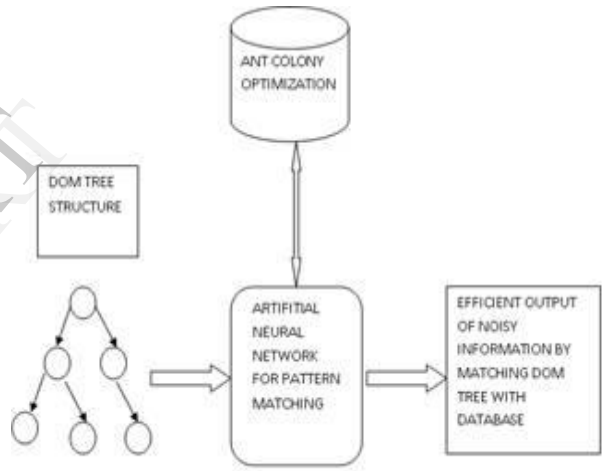

Figure5: Information matching using DOM and Ant Colony Optimization

V. CONCLUSION

Organizing and removing noise from web pages will get better on correctness of search results as well as explore speed, and may advantage web page association purpose (eg. keyword based search engine). For removal of noise Dom tree construction is always feasible as it converts the complex page into simplified form. Ant Colony Optimization algorithm has many advantages which helps to store and retrieve better results from the database .Neural Network always gives better pattern matching so that the information got after matching as almost similar or same. We can conclude that our proposed method are feasible to clean noisy data from web pages of any web site and the information retrieved from it will be accurate and better.

ACKNOWLEDGMENT

The authors would like to thank various authors whose paper has helped us to develop new idea and create the proposed architecture. We would like to thank our institutions for encouraging us to write the paper and propose the solution.

REFERENCES

[1] "Neural Networking using Multiple Web Page Noise Removing Method" P.Siva Kumar, Dr. R.M.S Parvathi IJCST Vol. 3, Issue 1, Jan. - March 2012

[2] "Neural Networks In Data Mining" ,dr. Yashpal Singh, alok Singh Chauhan. Journal of Theoretical and Applied Information Technology.

[3] "Elimination of Noisy Information from Web Pages" ,Alpa K. Oza, Shailendra Mishra. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-1, March 2013.

[4] L. Yi, B. Liu, X. Li., "Eliminating Noisy Information in Web Pages for Data Mining", in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Washington, DC, USA, 2003.

[5] Y. Yang, H. J. Zhang, "HTML page analysis based on visual cues", In Proceedings of the Sixth International Conference on Document Analysis and Recognition, pp. 859– 864, Washington,DC,USA,2001.