International Journal of Advanced Research in Engineering and Technology (IJARET)
Volume 11, Issue 10, October 2020, pp.161-170, Article ID: IJARET_11_10_016
Available online at http://iaeme.com/Home/issue/IJARET?Volume=11&Issue=10
ISSN Print: 0976-6480 and ISSN Online: 0976-6499
DOI: 10.34218/IJARET.11.10.2020.016

© **IAEME** Publication **Scopus** Indexed

WEB USER PROFILE IMPROVISATION BY SAMPLING SITE STYLE TREE WITH DOM STRUCTURE AND NEURAL NETWORK

S. Ganeshmoorthy
Ph.D. Research scholar, Department of Computer Science, Sree Narayana
Guru College, K.G.Chavadi, Coimbatore, Tamil Nadu, India

Dr. R. Priya
Associate Professor & Head, Department of Computer Science, Sree Narayana
Guru College, K.G.Chavadi, Coimbatore, Tamil Nadu, India

ABSTRACT

In the present generation, the web domain is the most rising and open knowledge medium. The Web Domain consists of contents for numerous areas, including multimedia, organized, semi-structured and unstructured data, which are accessible on the web to users with knowledge that is relevant. But within a given application only part of the information is relevant, and the remainder of the information is regarded as noises. Webpage details provide code formatting, links for navigation, advertisements and so forth. This set of unwelcome noise with specific content on a web page allows it more challenging to retrieve and process the automatic details. For this, usable noise-free data must be extracted. In this research work, we introduce a technology focused on the observation method for noise removal. Noisy blocks typically include similar content and design types on a given website, while the key content blocks of the websites sometimes vary in their content or design styles. On this basis, the tree layout, called StyleTree (ST), is suggested to collect the existing types of design and page content of a specific web site. An ST for the domain that we call SiteStyleTree (SST) can be generated by sampling the pages of the website. This role is followed by a potential application of NeuralNetworks (NN) to obtain material knowledge in the combination of three frameworks classified with the DocumentObjectModel (DOM). The sort of neural network used to develop our method utilize the Back-Propagation (BP) method in NN. Data were obtained from various Web servers for training and research. To remove different noise variations on the internet, the classification effects of a BP-NN were used. Experiments prove that our way of extracting insightful content from these websites' webpages is applicable efficiently. The comparison of the proposed work with existing work Noise Web Data Learning (NWDL) is done by parameters noise classification and accuracy. Thus the proposed work produces a good level of accuracy.

Keywords: Noisy data, Site Style Tree, DOM, Neural Network, Back Propagation

Cite this Article: S. Ganeshmoorthy and Dr. R. Priya, Web User Profile Improvisation By Sampling Site Style Tree With Dom Structure And Neural Network, *International Journal of Advanced Research in Engineering and Technology (IJARET),* 11(10), 2020, pp.161-170
http://iaeme.com/Home/issue/IJARET?Volume=11&Issue=10

1. INTRODUCTION

As the Internet expanded rapidly, the WorldWideWeb (WWW) has become a common site to disseminate and gather data. The loaded hyperstructure is noticed by users as the development of the Network is huge. A fast and effective updating of incoming data and the extraction of only required data is increasingly an issue among database mining testing communities [1], thus it needs to produce results without redundancies from the database. One of the data mining on the internet is termed as Web Content Mining (WCM) mechanism that allows end-users to more efficiently reach the web over restricted platforms, such as Personal Digital Assistants (PDAs) and mobile phones, to gain more valuable data, details, and understanding from web page contents that have several applications. Diverse data like images, voice, video and texts are used in WCM. The three types of Web-based data are most significant they are hidden data, core data and redundant data. The data of the web site's purpose is termed as hidden data which are the details behind the codes, headers and footers in news based webpages. The exact purpose to be derived from a Web page is defined as core data, for example, the headlines were a main subject in the news based web pages. The redundant data is the replication of data in the web page, such as the advertisement and details of the headlines repeated in many places on the news website. The input data goes through three steps in a Web mining process to achieve the result: Pre-processing, Datamining and Post-processing. Pre-processing may involve the deletion of unnecessary attributes and the cleaning up of noisy results. The information on a website that doesn't pertain to the key contents of the page is defined as web noise [2]. These noises are such as advertising banners, window guides, photos, etc. The contents are subdivided into smaller semantically homogenous parts until utilizing the material of a web page. Based on their granularity, site noises may be classified in a few groups.

The valuable data, knowledge and details from web pages are mined, processed and incorporated in WCM. The quantity of internet knowledge is increasing exponentially and the quantity and scale continue to increase steadily. With a large number of web sites, we can not be shocked that web users are constantly searching for relevant details utilizing search engines and search utilities. Thus, mobile data mining becomes an important way to discover valuable details or web-based expertise. However, a great deal of noise or disturbance, such as banner publicity, navigation bars, copyright notices, etc, is also followed by valuable facts or expertise on the internet. The details on the web page are beneficial to an individual, but it hampers the information collection phase due to different forms of noise on this web page. Since DOM trees stay extremely editable, it can quickly be restored to a full website. The DOM tree is organized hierarchically and can be studied completely or partially with a broad variety of versatility. By scanning a website through a DOM tree, it can be monitored more effectively where the noise could be eliminated. A web page is the first step to process into many regions or sections. The first step is to create a framework or model. A web page is separated into sections by many techniques. An HTML document is described as a DOM tree [3] in the DOM-based segmentation approach. Based on this finding, we propose a method to collect the typical design types and the current content of the sites on one website, the layout of the tree is named as ST. Our framework then generates the SST framework for a new web page and partitions it into subtrees and it as given as input to the second process for a NN. An ST will be generated for the website we name the SST if we examine the page of the website. The rest of the paper is organized as follows the related works in Section 2, the methodologies with the existing and proposed model are briefed in Section 3, the result and discussion with comparison are detailed in Section 4 and finally, the article was concluded in Section 5.

2. RELATED WORK

In [4] the author claimed that noisy and meaningless items needed to be removed from the web pages which have a broad variety of applications, including web page graders, website clusters, proper search engine indexation, enhancement of search result quality and text description. Therefore, cleaning web pages is obligatory to boost the efficiency of knowledge retrieval for web data extraction. Instead of eliminating related text blocks from web pages, they concentrate on reducing diverse noise trends.

In [5] the author said that the Internet was the biggest knowledge repository in any region. The noise in the site has been noted as a web page includes details that are not of interest to the customer. Noise from web pages impedes productive mining of web data. It is also incredibly critical that sounds of this sort are avoided. Identification and exclusion of noisy data on a web page are a pre-processing of different applications such as web page classification or data extraction, view of Web pages on computers, PDA, cell phones, etc. Many methods utilize the DOM tree principle to reduce web pages noise. The insightful quality of a web page is of concern to several people. The core components of the web sites must also be isolated from other parts of the material. Web sites may be separated into blocks of the web page where the knowledge is gathered. The elimination of blocks of non-content from web sites decreases the capacity and indexing needs.

The web has grown over several years to become the greatest archive of human knowledge. Websites have details that degrade mining data efficiency. The above-mentioned category of knowledge is regarded as an internal or external web noise. Here it suggests a tag processing strategy for the elimination of internal noises from a web page. Evaluating the properties and meaning of tags will exclude redundant photos and connections from a web page. The filtering of tags describing noise, or changing the properties of tags, may delete noisy content. Inclusion of picture publicity and background photos, nonimportant connections, search panels, copyright material, etc, are the items on the site page that are deemed a noise by the proposed venture. Precise, retrieve and F-score quality are tested for the elimination of picture advertising. The web pages have displayed a strong compression ratio since noise deletion, suggesting a substantial reduction in load time. The scale of the source code on the web pages was also substantially reduced as a consequence of removing sound tags from web pages.

3. METHODOLOGIES

3.1. Existing System

3.1.1. Noise Web Data Learning (NWDL)

The NWDL is an algorithm based on Machine Learning techniques that use the User-Profile (UP) in a web environment for learning noise before being disabled. A core emphasis of this approach is the understanding, detection and removal of noise by taking into consideration the user's diverse behavior and changing site data. Noise reduction through the collected log from the website data is calculated based on what a customer is involved in and omitted. The interest of the user on a web page is also calculated according to the number of occasions the user visits the website, the amount of time they spent on this website, the latest page visits and the number of connections that are on the page they visit [6]. This approach functions to a certain degree to measure the user's interest in web data logs retrieved but lacks insufficient information that was given to show how noise is measured on a site through UP before deletion.

A UP includes a range of URLs of user interest. The design of a UP is focused on a certain web page that a consumer accesses with due regard to its value. A variety of sessions are used in the UP. This work learns the degree of user engagement on visited web sites, after building a UP, to decide helpful details on noise results. Different metrics, e.g. length, duration and scope of visits to a web page by the customer are considered to process the results. This approach acknowledges the need to locate valuable knowledge to discover the interest of consumers. This can be achieved by the compilation, review and evaluation of user log data through a UP. The interest of the consumer depends on the assumption that the period that website visitors is an indication of the degree of interest of the consumer. In one session, the time spent on a collection of web pages accessed by the consumer represents the user's interest. It also means that a person displays a greater interest in web sites with a higher frequency. While this study takes into account the page visits and the pace of the visit, it is impossible to quantify user interest levels depending on the visiting period and visiting pace. For example, the high volume of website visits might represent a customer who is trying to locate valuable information, or who needs to visit several pages before reaching the information of interested parties depending on website architecture. Therefore, more measurements such as visits depth and duration to a website segment are included in the proposed work to collect usage information before noise data is excluded.

3.1.2. Suffix Tree Clustering (STC)

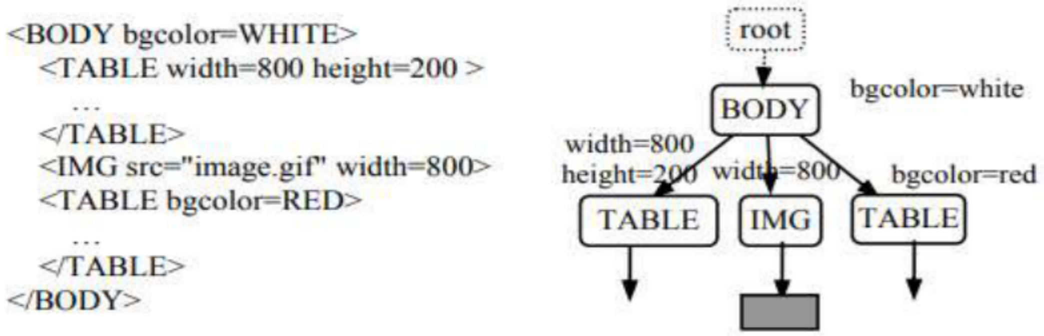
The STC is a Linear-Time method that works as a Clustering concept which focused on the phrase detection in the given text classes. A structured series of one or more words is a phrase in our sense [7]. To compile a document with a common-phrase we identify a base cluster. It operates by 3 conceptual stages: (1) the 'cleaning' text, (2) a suffix tree for recognition of the basic clusters, and (3) forming a cluster by combining the above two. In STC the tree structure forms as the text string of a document are translated by eliminating the suffixes and prefixes also by reducing plural to unique. The boundaries of the sentences are described as (HTMLTags and punctuations) and tokens of non-word like (Punctuation, HTMLTags and numbers) are eliminated. The exact text strings and markers are retained in the converted string from the beginning of each phrase to its exact location. This helps us to view the exact text to improve user readability until we recognize the main sentences in the transformed series.

3.2. Proposed System

The StyleTree (ST) a modern tree form is proposed to collect the exact text and typical templates (or display styles) of the webpage on a website. A calculation based on knowledge (or entropy) is often added to calculate the value of each entity node in the ST, which lets us reduce noise on a webpage. Then the framework generates an entrant web page with SST tree layout and separates them into sub-trees to get NN input data in the second position. Finally, for the webpage classification, here it uses a method BP-NN. The classification is based on the mixture, core data and group of noises.

3.2.1. Document Object Model (DOM)

The WWW-Consortium produced a model called DOM which works as an object-based framework that generates an XML and HTML document as a memory tree layout. The software uses the XML documents through the memory tree, it's a duplication of the layout of the code. The DOM helps users to cross and edit the XML document dynamically as well. It is not only for one particular Template tag but also a model for the entire text. A text as a tree is seen in the DOM. DOM trees are extremely convertible and can be used to restore a whole website quickly. The DOM tree is an HTML document model that's well established. There is no closed bracket for few tags of HTML. Some kinds of tags are closed by the following procedure, such as < LI > tag with the following tag as < /LI > tag is closed [8]. Every page of HTML is a DOM tree, with the tags being inner nodes, and the leaf nodes are informative texts, photos and hyperlinks. A section and its related DOM tree are shown in Figure 1. Any solid rectangle is a tag node in the DOM tree. The shaded box is the actual node contents, e.g. for the IMG Tag, "src = image.gif" indicates the original material. Note that our HTML webpage analysis starts with the BODY tag and all of the sections to be displayed are in the BODY area. The display properties of each node are also interconnected.



```
<BODY bgcolor=WHITE>
  <TABLE width=800 height=200 >
  ...
  </TABLE>
  <IMG src="image.gif" width=800>
  <TABLE bgcolor=RED>
  ...
  </TABLE>
</BODY>
```

Figure 1 DOM Structure

To render the analysis simple, we introduce a virtual root node with no attribute in the DOM tree as BODY's parent tag node. While the DOM tree must reflect a single HTML page's layout or design style, the total design style and content of a group of HTML pages can not be analyzed and are cleaned dependent on a specific DOM tree. In the cleaning process it takes the presentation style and actual web pages into consideration, thus, DOM trees are not enough. To process this we need a better system. This is a hard structure since our algorithm needs it to recognize similar types of webpages to remove noises [9]. For this, a new tree structure, StyleTree (ST), is implemented to compact a variety of linked Webpage's typical presentation types.

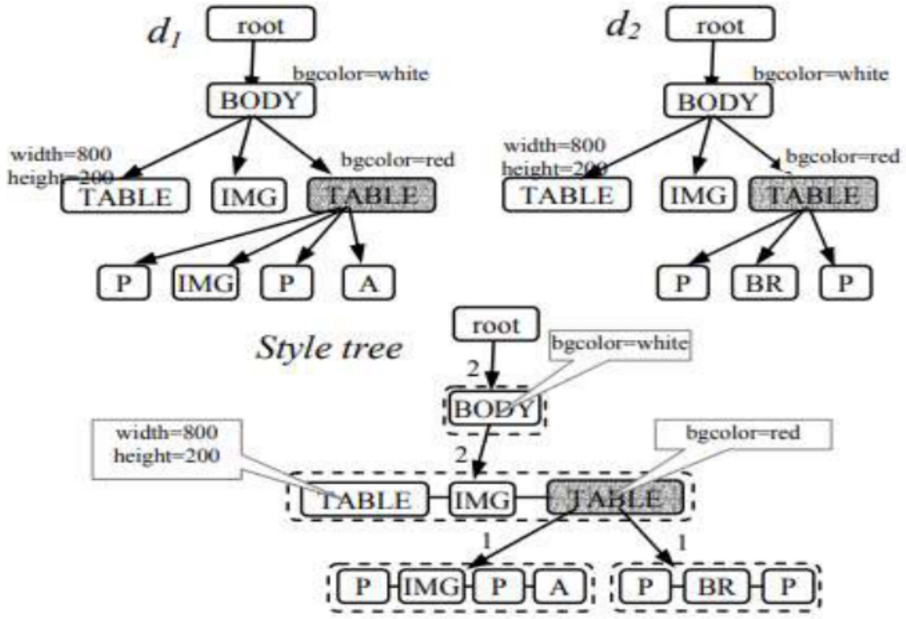


Figure 2 Different Style Nodes

Figure 2 shows a mixture of DOM trees d1 and d2 provides an example for ST. Here it notes all the tags in d1 have their matching tags in d2 except for the four tags (P, IMG, P and A) at the bottom stage. This will compress d1 and d2. Here it utilizes counts to determine how many pages are in a given type at a certain tree stage. Figure 1 reveals that the two pages begin with BODY, so the counts of the BODY is 2. The two pages under BODY both have the same TABLE-IMG-TABLE style-design. This complete series of tags is named a style-node (TABLE-IMG-TABLE), which was shown in Figure 2, embedded into a dashed-rectangle. At this stage, it reflects a certain type of design. Thus, in a DOM-Tree, a style-node is a set of tag-nodes. These tag-nodes are named element nodes in the ST so that they are distinct from tag nodes in the DOM-Tree. The style-node TABLE-IMG-TABLE for example comprises 3 element-nodes of the TABLE, IMG and TABLE. An element-node often includes knowledge that is subtly different from a tag-node inside a DOM-Tree. In Figure 2 it indicates the most divergence in the Tag TABLE, d1 and d2, that could be found in two separate style-nodes below the top in the ST. P-IMG-P-A and P-BR-P are the two style-nodes. It implies that we have two separate design types under the right node TABLE. These two style-nodes have a page count of 1. The ST simply displays the two DOM trees as compressed. It helps us to see what sections and which sections of DOM-Trees are similar.

We first search the HTML-Document syntax to evaluate a website because the majority of HTML webpages are not developed well. And then we use an HTML-parser to render a correction for the markup and a DOM-Tree creation [10]. The module breaks it into many sub-trees by the threshold, after constructing the DOM-Tree. Different Internet pages have varying interface and view types, so the width of the webpage tree differs based on the layout type. To select the right threshold level, the device wants to understand the upper limit of the DOM-Tree. The device then crosses the entire DOM-Tree to achieve the full DOM depth. We want to select the right level of the threshold for the training data collection, by defining the different levels of thresholds. The framework then uses these established series pairs to determine the acceptable threshold level for test results. Based on a linear regression study, the method determines the relationship between the peak and threshold limit levels. Regression is a mathematical analysis that measures the relationship between two variables [11]. RegressionAnalysis (RA) is often used to explain and investigate types of these correlations, which are the connections from the independent variables to the dependent variable. If the threshold level has been reached, the framework can identify any DOM nodes below the threshold level as noise and delete them before the grading phase starts. After breaking sub-trees, the input patterns for the NN classifier with Eq. 1 will convert into numerical representations.

$$Xi = Sn/Tn \tag{1}$$

Where Sn is the number of same leaf-nodes occurring in the sub-tree, Tn is the cumulative number of sub-tree leaf-nodes. More testing for the proposed method may be performed by translating a webpage into a DOM-Tree. After this may implement the BP-NN method to categorize three groups as data, noise and mixture (data and noise). Finally, we delete the webpage noise class and view the exact material derived on the HTML-page.

3.2.2. Site-Style Tree (SST)

A style-node (S) is a form or design of presentation with two components, defined by (Es, n), in which the element-nodes series is "Es" and the number of pages "n" for this specific style at this node-level. There are three components (TAG, Attr, Ss) of an E element-node consists: TAG is the name of the tag, e.g. "IMG" and "TABLE". The attribute set to display for TAG's such as "width = 100", "bgcolor = RED", etc. Under E the style-nodes set is denoted as Ss

To observe that there is an element-node in the DOM-Tree which corresponds to the tag-node, also by the merits to the collection of Ss child-style-nodes shown in Figure 3. By simplifying, the element-node is typically labeled with its "tagname" and a style-node with its "tag-names" sequence. It's pretty easy to create an ST (such as Site-StyleTree or SST) for a website. First, for each tab, we create a DOM-Tree and merge it in top-down style into an ST. In a unique "E" element-node in the ST, that has the corresponding tag-node "T" in the DOM-Tree, it then searches whether the child tag-nodes sequence "T" in the DOM-Tree is similar to the sequence of the element-nodes in the style-node "S" under "E". If it is "Yes", just raise the node S page counter to fuse the other nodes down the ST and the DOM-Tree. If it is "No", a new style -node will be generated in the ST below element-node "E" [12]. When translated to style-nodes and element-nodes of the ST, the sub-tree of tag-node "T" in the DOM-Tree is copied in the ST. The concept of noise in our work is based on the assumptions which follow as: (i) The element-node has many style types of presentation, it's highly significant as well as vice-versa. (ii) The element-node was more varied than the original substance, it's a highly significant element-node as well as vice-versa. In the assessment of the quality of an entity-node, all these essential values are used.

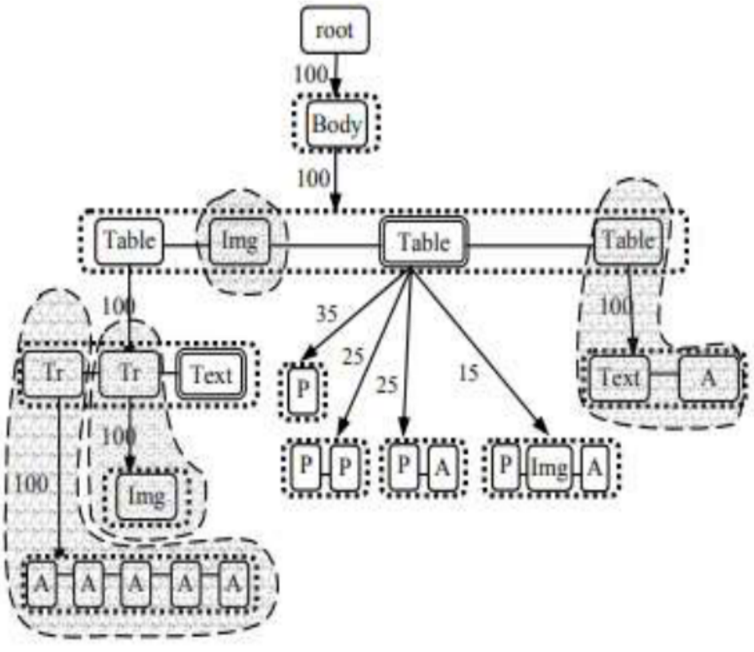


Figure 3 Structure of SST

The emphasis of the presentation is to recognize noise in standard styles whereas the value of information is to recognize certain key components of pages that could be portrayed with similar styles. Thus the importance of an entity-node is provided in the proposed method by comparing the importance of its appearance with the importance of the material. The more significant an element-node is in conjunction, the more often it is the essential material of sites. In Figure 3, for example, the SST shaded sections are more likely to be noise because they are extremely normal, fixed and therefore less significant in their presentation styles (including their original contents, which can not be seen in Figure 3). There are several child style-nodes in the double-lined table element-nodes, which demonstrate that the element-node is possibly significant. In other terms, it is more probable that the double-lined table comprises the principal material of the pages. The text element-nodes with dual lines are also important, as their contents are varied, while their type of presentation is set. Let the SST be the ST generated for all website sections.

3.2.3. Identification of The Content Using Neural-Network

The NN model can be considered to solve problems efficiently while comparing with the traditional algorithms as a successful problem-solving process. If the feedback is something it had ever seen before, it responds identically to the one that fits the nearest data template for training. It receives the values as the input, computes and transfers to the subsequent layer of each node in the NN model architecture. This method is carried out by weight that is the power of the relation between two nodes. A NN is a system consisting of many basic items or connections known as neurons. These components still function concurrently. The NN's role is primarily defined by the neuronal relation. These neurons are related and each relation is modified by weight values. Learning is named as it learns dynamically in the way of weight updating. The "p" input of neuron is connected to weight "w" and the "b" biases are scalar. Equation 2 forms a reference for the second part of the function which is transfer.

$$n=wp+b \tag{2}$$

The neuron output is the transmitting mechanism output. It equated as follows:

$$a = f\,(wp+b) \tag{3}$$

Where "f" is a Transfer-function that takes "n" arguments and generates an "a" outcome. By changing its parameters, NN may display the required or interested actions. This implies that NN could be trained based on a certain role by changing the weight or bias parameters, or the network may change this parameter for optimal performance. The "p" input to the neuron may be raised to "R" elements and weight is compounded with each input. Its sum is only (W●P) the product of the "W" matrix and the "P" vector. The "n" argument was neuron's Transfer-function input is given as:

$$n = w1,1\ p1 + w1,2\ p2 + w1,3\ p3 + \ldots + w1,Rp\ R + b \tag{4}$$

The multi-layer feed-forward BP-NN is one of the most widely deployed NN. It comes under the "Classification and Prediction Networks" group. This particular form of NN is used to construct our framework. Our model is constructed of two layers in which the neurons are organized logically. The last-layer is the output-layer and only one layer was hidden. Several pages of various web sites used as a data collection were randomly chosen to train the model. This analysis aims to solve a multi-class issue in which it not only separates noise patterns from web pages but also defines data and mixing patterns [13]. Both the NN deployed had 15 neurons and 2 noise-pattern neurons output [0,1], data-pattern [1,0] and mixture-pattern [1,1]. The BP is well suited for training and elaborately applicable method. The basic sigmoid activation method for both layers was found to be appropriate for the model of regional classification of the web page. The NN classification outcome is used to eliminate separate noise patterns. Not only the accuracy of the classification result but also the threshold level of the DOM-Tree may decrease due to noisy information in web pages. Hence the BP-NN will be very effective in the reduction of noises in the webpages.

4. RESULTS AND DISCUSSION

The findings are focused on the classification of noises and level of noise accuracy and the study being implemented and compared with the NWDL-STC method.

Unlike existing algorithms, the reduction of noise has been eliminated in these decades and is solely represented by shifts in user demand over the years. Although an individual uses the class over time as existing noise knowledge is used to delete noise from site data and only user details to be excluded. The frequent interest of the user can not be determined by the amount and length of visits to a website alone. A user could be more motivated over a certain amount of time. Consequently, the amount and time of the operation are no longer easily noticeable for users while site recordings are being cast off and are shown in Table 1 and Figure 4.

Table 1 Noise Classification Ratio

S. No	Methods	Interest	Noise
1	NWDL+STC	146	82
2	Proposed System	146	96

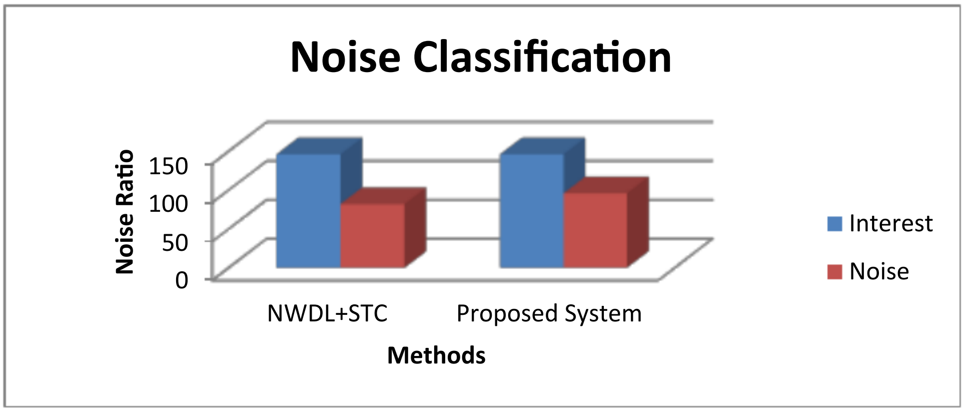


Figure 4 Noise Classification

The accuracy of the noise-reduction not only relies on the accurate classification outcome of the NN but also the thorough division of sub-trees. Multiple sub-trees can be categorized into framework heuristics because websites have complicated and diverse architectures with a range of websites. The three places on the webpage may be categorized for chosen websites with percentage values as indicated by Tables 2 and Figure 5. The percentage of noise and datatype or level of mixture in the multiple webpages of chosen websites is displayed. By incorporating the BP-NN for DOM-Tree the different noise-patterns are effectively excluded and detailed site data retrieved with an accuracy rate of 97%.

Table 2 Accuracy

S. No	Methods	Quality Accuracy (%)
1	NWDL+STC	90
2	Proposed System	97

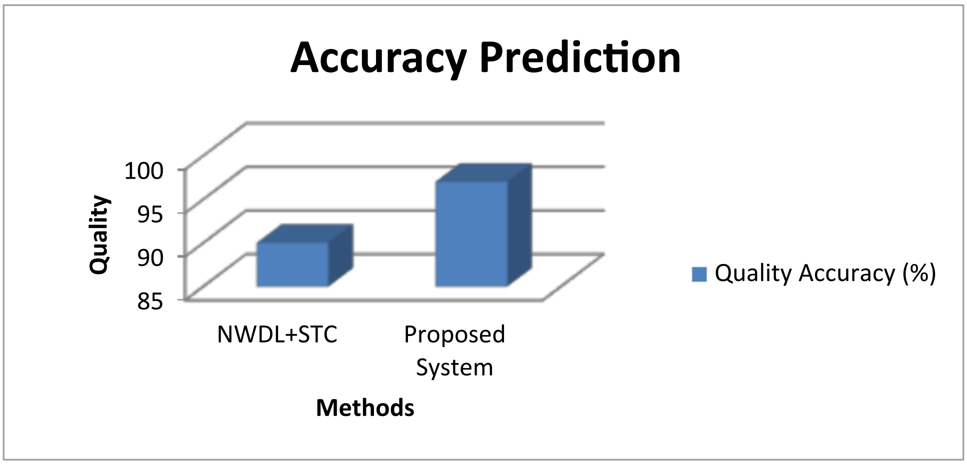


Figure 5 Accuracy

5. CONCLUSION

It has been measured the amount of noise class sub-trees that are observed and extracted on each webpage of selected websites to assess the system efficiency. The proportion of noise-trees determined by the total value of sub-trees collected by the system is determined for each area (noise, data and mixture). Three classes were overcome by the proposed method and the noise class was eliminated. Here it not only build a method for classifying three groups on the web page, but also for eliminating the noise class based on neural network classification findings. Both noise classes on the identified domains are essentially removed by the proposed framework. It provides the proper grading outcome not only by the neural network but also based on the threshold level to separate DOM sub-trees, that improvizes the accuracy of this System's noise removal.

REFERENCES

[1] Dutta, A., Paria, S., Golui, T., & Kole, D. K. (2014). Noise Elimination from Web Page Based on Regular Expressions for Web Content Mining. Smart Innovation, Systems and Technologies Advanced Computing, Networking and Informatics- Volume 1, 545-554. doi:10.1007/978-3-319-07353-8_63

[2] Yi, L., Liu, B., & Li, X. (2003). Eliminating noisy information in Web pages for data mining. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '03. doi:10.1145/956750.956785

[3] Yang, Y., & Zhang, H. (2001). HTML page analysis based on visual cues. Proceedings of Sixth International Conference on Document Analysis and Recognition. doi:10.1109/icdar.2001.953909

[4] Uma, R., & Latha, B. (2018). Noise elimination from web pages for efficacious information retrieval. Cluster Computing, 22(S6), 14583-14602. doi:10.1007/s10586-018-2366-x

[5] Arya, C., & Dwivedi, S. K. (2018). Content extraction from news web pages using tag tree. International Journal of Autonomic Computing, 3(1), 34. doi:10.1504/ijac.2018.092548

[6] Dias, S., & Gadge, J. (2014). Identifying Informative Web Content Blocks using Web Page Segmentation. International Journal of Applied Information Systems, 7(1), 37-41. doi:10.5120/ijais14-451129

[7] Bar-Yossef, Z., & Rajagopalan, S. (2002). Template detection via data mining and its applications. Proceedings of the Eleventh International Conference on World Wide Web - WWW '02. doi:10.1145/511446.511522

[8] Garg, A., & Kaur, B. (2014). Enhancing Performance of Web Page by Removing Noises using LRU. International Journal of Computer Applications, 103(6), 23-27. doi:10.5120/18079-8632

[9] Roychowdhury, S., & Pedrycz, W. (2018). Automatic Discovery of Clusters by Removing Noisy Data. International Journal of Intelligent Systems, 33(9), 1777-1797. doi:10.1002/int.21904

[10] Shah, H., Rezaei, M., & Fränti, P. (2019). DOM-based keyword extraction from web pages. Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing - AIIPCC '19. doi:10.1145/3371425.3371495

[11] Kumar, P., & Dadheech, P. (2019). Performance of Fuzzy Filter and Mean Filter for Removing Gaussian Noise. International Journal of Computer Applications, 182(38), 29-35. doi:10.5120/ijca2019918399

[12] Sivakumar, P. (2015). Effectual Web Content Mining using Noise Removal from Web Pages. Wireless Personal Communications, 84(1), 99-121. doi:10.1007/s11277-015-2596-7

[13] Aghamaleki, J. A., & Baharlou, S. M. (2018). Transfer learning approach for classification and noise reduction on noisy web data. Expert Systems with Applications, 105, 221-232. doi:10.1016/j.eswa.2018.03.042