International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Developing Decision Tree for Measuring Impact of Errors in Web Page Content using Data Mining Techniques

J Hari Narayana[1], G. Sreedhar[2]

[1]Research Scholar, Department of Computer Science, National Sanskrit University, Tirupati

[2]Department of Computer Science, National Sanskrit University, Tirupati

Abstract: ***The research paper proposes a methodology to develop decision tree for measuring impact of errors in web page content withheld of data mining techniques. The website with poorly designed web pages no longer serves the purpose of internet users. The methodology is very much comprehensive and it covers all aspects of web page content for hundred percent satisfaction of end user.***

Keywords: web page content, web site errors, decision tree, impact of error

1. Introduction

Web designing is the process of creating a website and made it available to online viewers. Website is the collection of web pages, documents and multimedia over the internet. In present society designing a quality website is the challenging task to the web designers. In order to provide hundred percent intended services of the website to the end users the website design process must be thoroughly verified in web content, web structure and web usage. Web mining techniques are used to provide quality web design for end user as well as for web developers. Web mining is the use of data mining techniques to extract knowledge from web data. Web mining is performed on three areas of web sites. They are web content mining, web structure mining and web usage mining. Web content mining is the process of mining, extracting useful information and knowledge from web page content. Web page content can be structured, semi structure or unstructured collection of data.

2. Related Work

In order to website design the web developer must concentrate on web page content development. The World Wide Web Consortium (W3C) defines a set of guidelines for quality web content. The W3C provide various online tools to identify the errors of web page content during validation process of web pages of website. The W3C tool HTML validator accepts the home page of website and list the all errors of website based on well defined guidelines. The errors which are listed in validation process are taken into consideration in evaluating their impact on display of web page content in the proposed research work.

3. Methodology

A comprehensive methodology is developed based on observations previous research scholars' contribution that encompasses all aspects that need to be concentrated in designing web page content. The web page content consists of collection of text, images, tables, hyperlinks and multimedia. The W3C HTML validator produces the list of errors related to web page content. The impact of errors of web page content need to be verified by the web developer so that end user is able view the all contents of website or not. The methodology consists of following steps.

a) Start validation using W3C HTML Validator for intended website.
b) Identify all errors of web page using W3C guidelines by the validator.
c) Calculate the standard deviation of all identified errors.
d) Develop a decision tree to measure the impact of errors on web page content.
e) If the impact of errors is 'Low' then the contents of the web page are visible with minor impact on display of web page content.
f) If the impact of errors is 'High' then all the content of web page may not be viewed by the end user.

4. Decision Tree Generation and Implementation:

A decision tree is flowchart like structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label. The top most node of tree is the root of the decision tree. The construction of decision tree classifiers does not require any domain knowledge of parameter setting and therefore is appropriate for exploratory knowledge discovery. The decision trees can handle high dimensional data and in general decision tree classifiers have good accuracy. The decision tree for proposed research work demonstrates classification of various errors of web page content. The decision tree is generated using random tree generation algorithm. The random tree generation algorithm covers all classified errors and hence correctly judges their impact on web page content. The errors that are included in measuring impact of web page content are Table Tag Errors (TTE), Form Tag Errors (FTE), Style Tag Errors (StTE), Frame Tag Errors (FrTE) and Document Type Declaration Errors (DTDER). The screen short of all types of errors considered in decision tree generation is shown in the figure 1.

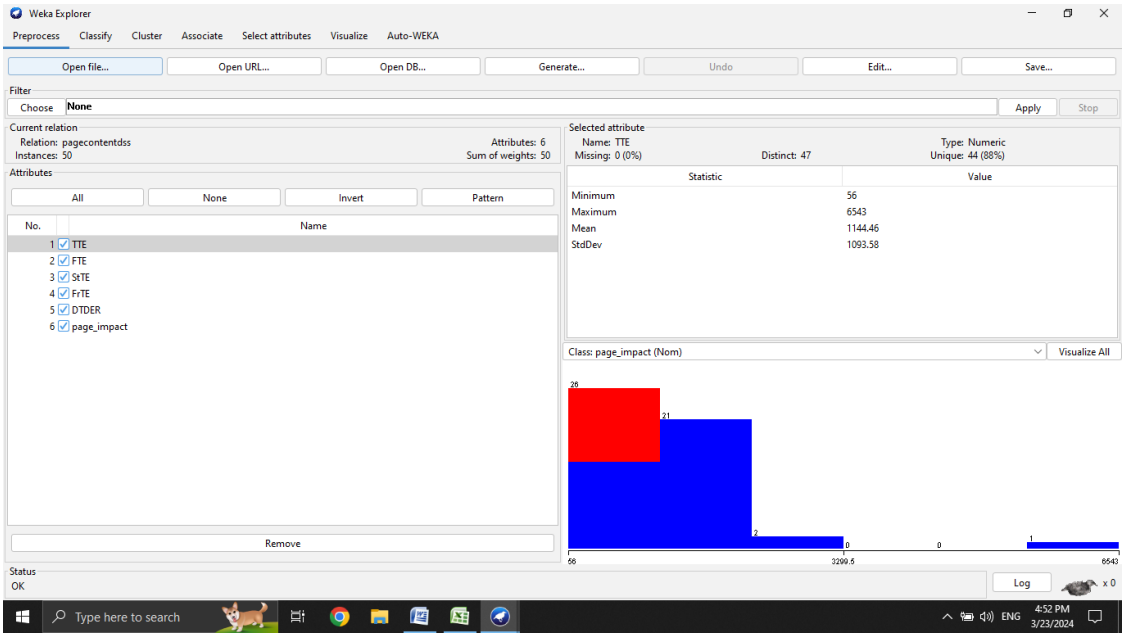

Figure 1: Preprocessing of Web page content errors

The decision tree generated by random generation algorithm correctly classifies all types of errors with 100% accuracy. Also decision tree identifies positive tuples and negative tuples with clear classification. The True Positive Rate (TP Rate) and False Positive Rate (FP Rate) of each class determine the classification accuracy measure. The confusion matrix is useful tool to recognize various types of classes. The confusion matrix for classification errors is a 'm X m' matrix which identifies impact of errors on web page content. The screen shot of classifier output of various errors is shown in the figure 2.

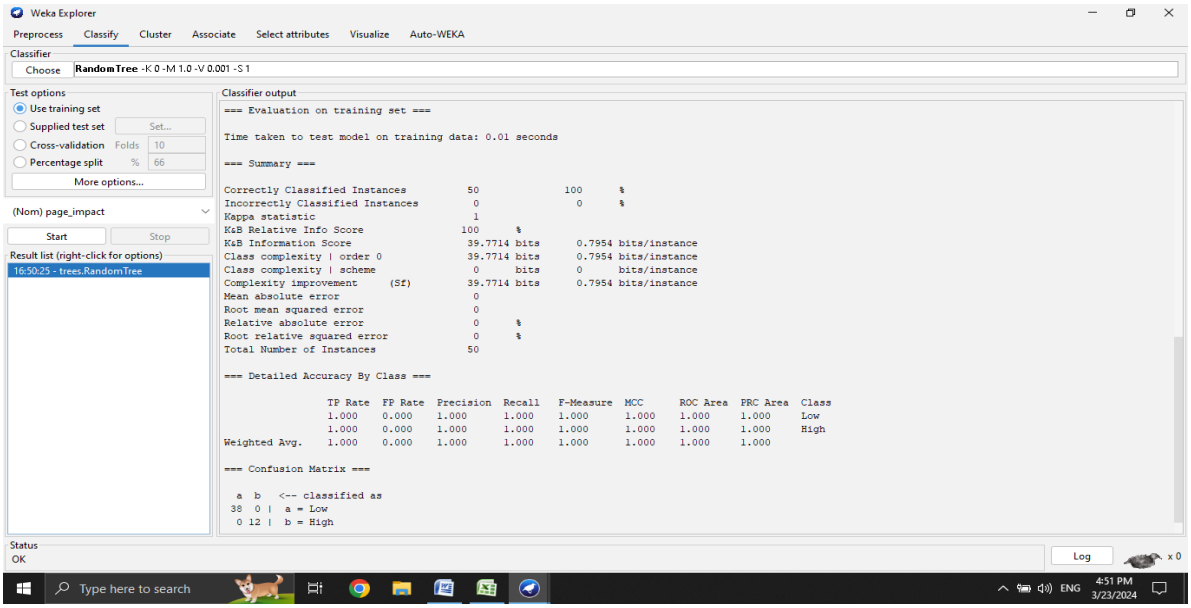

Figure 2: Outcome of various measures based on Random tree generation

The decision tree using random generation algorithm is shown the figure 3. The leaf node of decision tree shows the impact (either low or high) based on number of errors of each type of page content errors. The Form Tag Errors (FTE) is the root of the decision tree and based on its value further generation of branches in decision tree is classified. The combination of number of error sets {FTE, TTE, FrTE}, {FTE, TTE, FrTE, StTE} and {FTE, TTE, FrTE, DTDER} produce the high impact on web page content and other combinations of errors produce only low impact on web page content. The combination of errors and their impact are shown in the table 1.

Table 1: Web page errors and their impact

S. no	Web page errors combination sets	Impact on web page
1	{FTE}	Low
2	{FTE, TTE}	Low
3	{FTE, TTE, FrTE}	High
4	{FTE, TTE, FrTE, StTE}	Low
5	{FTE, TTE, FrTE, StTE}	High
6	{FTE, TTE, FrTE, DTDER}	Low
7	{FTE, TTE, FrTE, DTDER}	High

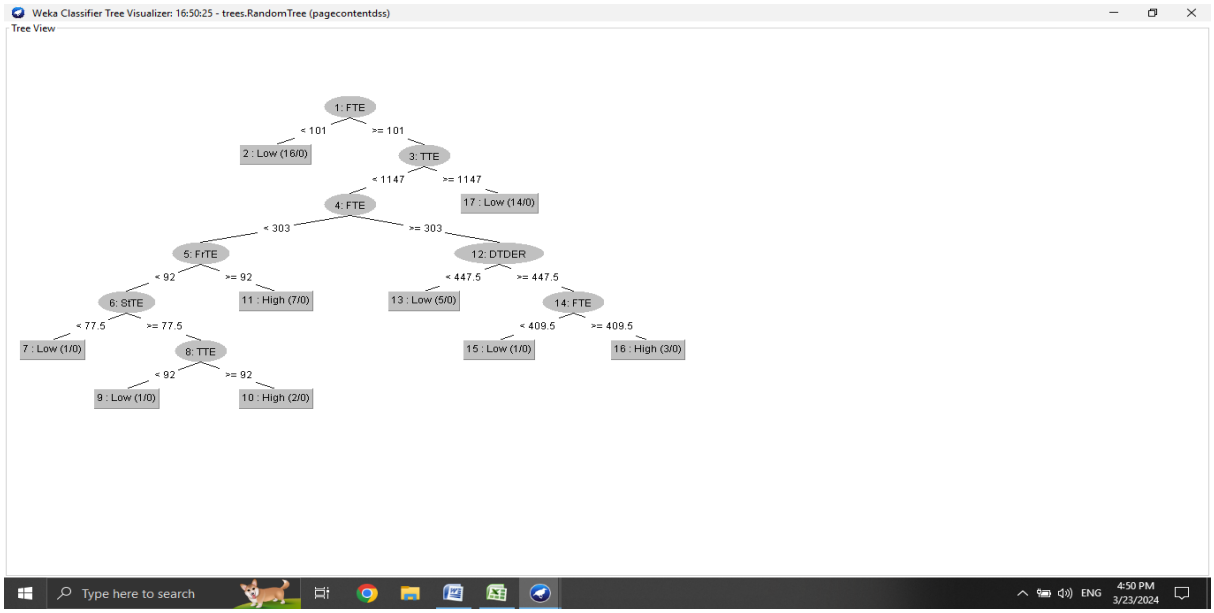

Figure 3: Decision tree of various errors of web page content with impact values

5. Conclusion

The main objective of research paper is to analyze various of types of errors related to web page content using decision tree. In this research article various data mining techniques like preprocessing, decision tree, random tree generation algorithm are used in order to know the impact of errors on web page content. In this paper an effort is made to know the impact of errors on web page content so that web developer and end user the can know the status of website pages. The decision tree can judge efficiency on display of web page content based on various combination of errors related to web page content.

References

[1] Signore., 2005. A Comprehensive model for websites quality. In: Proceedings of the seventh IEEE International symposium on Website Evolution (WSE'05).
[2] Bright, P., 2014. HTML5 specification finalized, squabbling, overspecs continues. From https: //arstechnicacom/information - technology/2014/10.
[3] G. Sreedhar, 2014. Analyzing download time performance of university websites in India. In: Internation Journal of web science and engineering. Vol.1. No - 1 (2014).
[4] G. Sreedhar., 2018. Improving usability of website design using W3C guidelines. In: Encyclopedia of Information science and technology, fourth edition.
[5] Layla Hasan, Enad Abuelrub., 2011. Assessing the quality of web sites. In: Applied
[6] computing and Informatics (2011), 9, 11 - 29.
[7] Alejandro, Rafael et. al, Website Quality: An analysis of Scientific Production", Interactive content and creation in multimedia information communications: audiences designs, systems and styles, September 2020.
[8] Semeradova, Weinlich et. al, "Looking for the definition of website quality", In: Semeradova, Tereza, Weinlichj, Peter, Website quality and shopping behaviour: Quantitative and Qualitative evidence. Springer Nature, PP.5 - 27, ISBN: 978 3 030 44439 6.2020
[9] www.validator. w3. org Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.
[10] Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD - ROM.
[11] Gordon S. Linoff and Michael J. A. Berry, Data Mining Techniques: for Marketing, Sales and Customer Relationship management. Third Edition, Wiley Publishing, USA, 2011.