Semantic-based Web Page Classification System Using Enhanced C4.5

Hnin Pwint Myu Wai[*], Nyo Nyo Yee[**], Nandar Win Min[**]

[*] Faculty of Information Science, University of Computer Studies Banmaw, Myanmar
[**] Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City),Myanmar

DOI: 10.29322/IJSRP.12.08.2022.p12832
http://dx.doi.org/10.29322/IJSRP.12.08.2022.p12832

Paper Received Date: 19th July 2022
Paper Acceptance Date: 07th August 2022
Paper Publication Date: 16th August 2022

***Abstract*-** Web is the largest collection of electronically accessible documents which make the richest source of information in the world. Classification of Web page is essential to many tasks in Web information retrieval such as maintaining Web directories, topic-specific Web link analysis and focused crawling. Traditional Web page classification ignores semantic relationships between keywords and Web pages. To effectively handle the semantic challenge, Semantic-based Web Page Classification System is presented. Moreover, a lot of research focuses to improve the existing classification algorithms used to classify web pages and most of the research is based on web page content. However, there is little focus on exploring new features that can be used to classify web pages based on not only content but also title and link of the web pages. For semantic, ontology is used to store each concept of each word and define links between different types of semantic knowledge. For classification, the Enhanced C4.5 decision tree classifier is applied. HTML pages from the computer science domain is tested to demonstrate its efficiency.

***Index Terms*-** Ontology, Classification, Enhanced C4.5, Semantic

I. INTRODUCTION

With the extreme growth of Web based information, Web page Classification process becomes one of the major challenges in organizing and maintaining the enormous collection of Web pages. Web page classification is useful for managing and extracting relevant information from Web content. Classification of Web pages can also help to improve the quality of Web search. The process of assigning a Web page to one or more predefined category labels is known as classifying/categorizing.

As the pace of Internet usage has rapidly expanded, there has been a massive increase in interest in the automatic classification of Web pages into specified categories. Several machine learning algorithms have been successfully used in the past by the scientific community. They include Neural Networks, Naive Bayes, Support Vector Machines (SVM) and k-Nearest Neighbors (KNN). Traditional algorithms learn the features of categories from a series of classified documents and use the classifier to classify texts into predetermined categories. However, these machine learning methods have some drawbacks: (1) In order to train classifier, human must collect large number of training text term, the process is very laborious. If the predefined categories changed, these methods must collect a new set of training text terms. (2) Most of these traditional methods haven't considered the semantic relations between words. So, it is difficult to improve the accuracy of these classification methods.

In order to solve these problems, this system is proposed as the Ontology-based Web Page Classification System by using Enhanced C4.5 decision tree classifier. This system uses the ontology to search the semantics for each feature in each Web page. Semantic-based Classification method considers the semantic relations between features extracted from Web pages, thus it improve accuracy of the classification method.

The rest of the paper is organized as follows: related work is described in section II. Web page classification is described in section III. Pre-processing for Web page is expressed in section IV. Decision tree classifier, C4.5 classifier and ontology are described in section V, VI and VII respectively. The proposed system design is detailed in section. Sections VIII and IX describe experimental results, respectively. Finally, conclusion is given in section X.

II. RELATED WORK

In 2005, M. Song, S. Lim and D. Kang [12] suggested an automated method for document classification using ontology. Using at least two predefined categories per given document characteristic, ontology-based document classification entails selecting document features that most accurately reflect Web documents and classifying them into the most appropriate categories after studying their

This publication is licensed under Creative Commons Attribution CC BY.

contents. In this approach, Web pages are categorized in real time using comparable calculations between the terminology information taken from Web pages and ontology categories, rather than experimental data or a learning process. As a result, the meanings and relationships specific to each document are determined, resulting in a more accurate document classification.
In 2008, the automatic document classifier system based on the Naive Bayes classifier and ontology was described by Y. Chang and H. Huang [10]. The main concept is to first establish a keyword synonymous table by experts for narrowing down the range and getting the consistency of keywords. The formal concept analysis is then used for establishing knowledge ontology through the complex categories and attributes relation. Finally, the ontology is applied to a Naïve Bayes Classifier to get the automatic document classifier system. In this system, 319 documents divided into 11 categories are used to assess the effectiveness of classification, where 224 and 95 documents are the training and testing documents respectively.
In 2012, W. K. Ong, J. L. Hong and F. Fauzi [11] presented a novel and fast ontological-based Web page classification technique to classify a Web page with high accuracy. To speed up our system, we use a segmentation technique that utilizes visual boundary of a region and matches keywords within the region instead of the entire Web page. They used a fast clustering technique to match keywords and label the page based on the nearest match. Experiment results show that their system is accurate in Web page classification.

III. WEB PAGE CLASSIFICATION

The technique of assigning one or more specified categories (topics) to Web pages based on their content is known as web page classification. It can filter Web pages and send them to topic-specific processing methods such as data extraction and machine translation. The automatic Web page classification consists of two phases that are learning phase and classification phase. In the learning phase, users define categories (topics) in which they are interested (their information need) by giving sample documents (training examples) for each of these categories. In the classification phase, new (previously unseen) documents can be given to the classifier which returns a topic. Machine learning, statistical pattern recognition, or neural network approaches are used to construct classifiers automatically [3].
There are three types of automatic Web page classification tasks: supervised Web page classification, where an external mechanism (such as human feedback) provides information on the correct classification of Web pages, unsupervised Web page classification, where the classification must be done entirely without reference to external information, and semi-supervised Web page classification, where the external mechanism labels parts of the Web pages [4].

IV. PRE-PROCESSING FOR WEB PAGES

Some preprocessing procedures are frequently performed before the Web pages in a collection are used for classification. For traditional text documents (no HTML tags), the tasks are stopword removal, stemming, and handling of digits, hyphens, punctuations, and cases of letters. For Web pages, additional tasks such as HTML tag removal and identification of main content blocks also require careful considerations. Stopwords are frequently occurring and insignificant words in a language that help construct sentences but do not represent any content of the documents. Common stopwords in English include a, about, an, are, as, at, be, by, for, from, how, in, is, of, on, or, that, the, these, this, to, was, what, when, where, who, will and with. Such words should be removed before Web pages are indexed and stored [9].

V. DECISION TREE CLASSIFIER

In data mining, a decision tree is a rapid and efficient classification technique that is commonly used for decision making. Decision trees are supervised methods for determining the link between input and target attributes, which is represented in a model structure. It examines the object's attribute values and chooses the best attribute as the root node, then uses that attribute to decide the tree branch's leaf nodes from the root down.
The majority of decision tree classifiers work in two stages: tree-growing (or constructing) and tree-pruning. The tree is constructed from the top down. The tree is recursively partitioned during this phase until all of the data elements are assigned to the same class label. The full-grown tree is pruned back in the tree pruning phase to minimize overfitting and improve the tree's accuracy from the ground up. There are two type of decision trees: tree structure (hierarchical structure) and rules structure (if-then statement) [5].

A. Types of Decision Tree Classifier

A decision tree is a graphical depiction of a tree with conditions connected with the nodes that allows a new instance to be classified into a preset set of classes. There are many types of decision tree classifier. These are CHAID (CHi-squared Automatic Interaction Detector), CART (Classification and Regression Tree), C4.5, C5.0 and Hunt's Algorithm.

B. Advantages of Decision Tree Classifier

Advantages of decision tree classifier are as follows:
- Decision trees are straightforward to comprehend and analyse.
- They only need a little amount of data and can handle both numerical and category data.
- Statistical tests can be used to validate a model.

This publication is licensed under Creative Commons Attribution CC BY.

- They are strong in nature, thus they perform well even if the underlying model from which the data were derived violates some of their assumptions.
- Decision trees can handle vast amounts of data in a short amount of time.
- Personal computers can evaluate large amounts of data in a short amount of time, allowing stakeholders to make decisions based on the results.
- Tree performance is unaffected by nonlinear interactions between parameters.
- The best thing about employing trees for analytics is that they're simple to understand and communicate to executives. [7].

C. Disadvantages of Decision Tree Classifier

Disadvantages of decision tree classifier are as follows:

- A slight change in the data can result in a significant change in the decision tree's structure, causing instability.
- When compared to other algorithms, a decision tree's calculation might be somewhat complex at times.
- The training period for a decision tree is typically longer.
- Decision tree training is fairly expensive due to the increased complexity and time required.
- For using regression and predicting continuous values, the Decision Tree approach is insufficient.

VI. ENHANCED C4.5 CLASSIFIER

C4.5 algorithm is an upgraded version of ID3. Instead of using gain as a splitting criteria in the tree development phase, C4.5 algorithm employs gain ratio as a splitting criteria. Both continuous and discrete properties are handled by this algorithm. In Original C4.5 Algorithm, the semantic logic is not considered to choose the root node (best attribute). In the proposed Enhanced C4.5, the root node is chosen by calculating normalized information gain that is based on both the original and semantic classes. The Enhanced C4.5 Decision Tree Algorithm is described in the following.

Algorithm: Enhanced C4.5 Decision Tree

Input: An attribute-valued dataset D

1. Tree = { }
2. if D is “pure” OR other stopping criteria met then
3. Terminate
4. end if
5. for all attribute a € D do
6. Compute gain ratio if we split on a based on original class level
7. Next, Compute gain ratio about a based on semantic class level
8. Combine each gain ratio result that is obtained by calculating each attribute value based on original class level and semantic class level
9. end for
10. abest = Best attribute according to above computed criteria
11. Tree = Create a decision node that tests abest in the root
12. Dv = Induced sub-datasets from D based on a best
13. for all Dv do
14. Treev = C4.5(Dv)
15. Attach Treev to the corresponding branch of Tree
16. end for
17. return Tree

At each node in the tree, the normalized information gain is used to choose the test attribute. This method is as follows:

$$Info(D) = -\sum_{i=1}^{m} P_i Log_2(P_i) \tag{1}$$

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Info}(D_j) \tag{2}$$

$$\text{SplitInfo}\ (A) = -\sum_{i=1}^{m} \frac{|Ci|}{C} \log 2 \frac{|Ci|}{C} \tag{3}$$

$$\text{Gain}\ (A) = \text{Info}(D) - \text{Info}_A(D) \tag{4}$$

This publication is licensed under Creative Commons Attribution CC BY.

$$\text{Gainratio}\ (A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(\ A)} \quad (5)$$

P_i is the probability that an arbitrary tuple in partition D. Info(D) is the average amount of information needed to identify the class label of a tuple in D. $|D_j|/|D|$ acts as the weight of the j^{th} partition. $Info_A(D)$ is the expected information required to classify a tuple from D based on the partitioning by A. C_i is the objects in class C that have value A of A_i. SplitInfo(A) is the information due to the split of class C on the basis of the value of the categorical attribute A. The attribute A with the highest information gain, Gain ratio (A), is chosen as the splitting attribute at Node N [8].

VII. ONTOLOGY

The core of semantic Web is ontology, which is used to explicitly represent conceptualizations. When combining numerous data sources, ontology resolves semantic issues. As a result, ontology is used as a common tool for data integration in many Web applications. Ontology also maintains information about the various data sources' concepts and relationships [1]. The semantic Web's ontology engineering is primarily supported by languages like RDF (Resource Description Framework), RDFS, and OWL (Web Ontology Language). Ontological frameworks are typically created using manual or semi-automated approaches, which necessitate the skills of developers and specialists. Ontologies are divided into two categories: those that are used to explicitly capture "static knowledge" about a domain, and those that provide a logical point of view about the domain knowledge (the second) (problem solving knowledge) [2].

VIII. PROPOSED SYSTEM DESIGN

The proposed system design is shown in Figure 1. The system is divided into two parts: training and testing. In Training, the user inputted Web pages are pre-processed and then extracts features from ontology. Attributes are collected to create the decision tree by pre-processing the web pages and then thresholds values for these attributes are observed by changing different values. In feature extraction, the proposed system extracts features not only from the content but also from the title and link of the web pages. Title of the web pages can be supported more related word for the classification. Therefore, the higher weight value to the title word of the web pages are assigned. Then, Enhanced C4.5 algorithm is used to produce decision rules to perform the classification process. In Testing, the user must input the new Web page that is needed to know category. Before classification process, this system performs preprocessing that includes tokenization and stopword removal. This system assigns the category according to the decision rules derived from Enhanced C4.5 decision tree classifier. Finally, this system produces the category as a result.

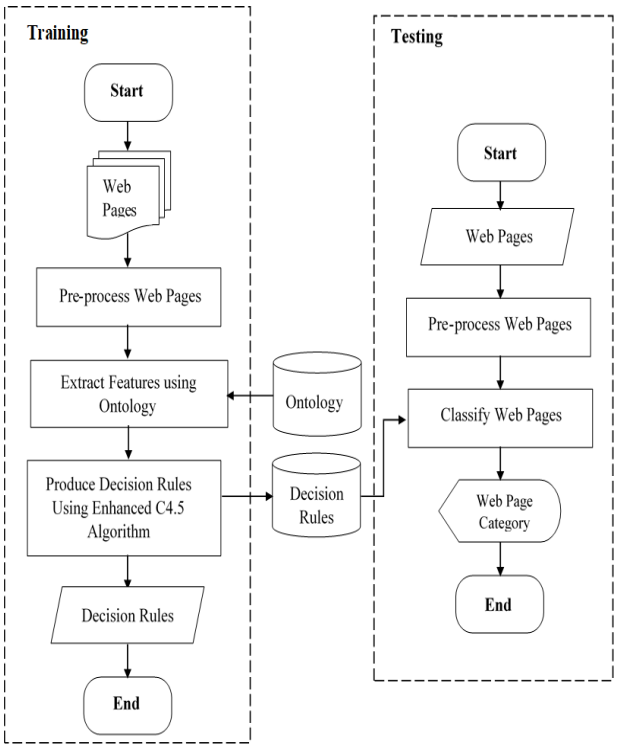

Fig. 1. Proposed System Design

IX. EXPERIMENTAL RESULTES

This publication is licensed under Creative Commons Attribution CC BY.

Most Web page classifications used only body contents of Web pages but this system not only used body contents but also title of Web Pages and link for the classification step to get the better accuracy. In this system, confusion matrix is used to measure the accuracy of the system. A confusion matrix can be used to precisely assess a classifier's potential. To measure the accuracy of the Ontology based Web Page Classification System, the data set is separated into two sets, called the training set and the testing set. 320 training Web pages and 80 testing Web pages about computer science domain are used to test this system. The classification performance of classifier is evaluated by the formula:

$$\frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Where, TP = True positive rate,

FP = False positive rate,

TN = True negative rate,

FN = False negative rate

TABLE I
ANALYSIS RESULTS WITH VARIOUS THRESHOLD VALUES ON BTL AND BODY USING ENHANCED C4.5 CLASSIFIER

Threshold Values	Body + Title + Link (BTL)	Body
Threshold = 3	0.875	0.675
Threshold = 5	0.875	0.575
Threshold = 7	0.750	0.575

Analysis results with various threshold values on BTL (Body + Title + Link) and Body are shown in Table 1.

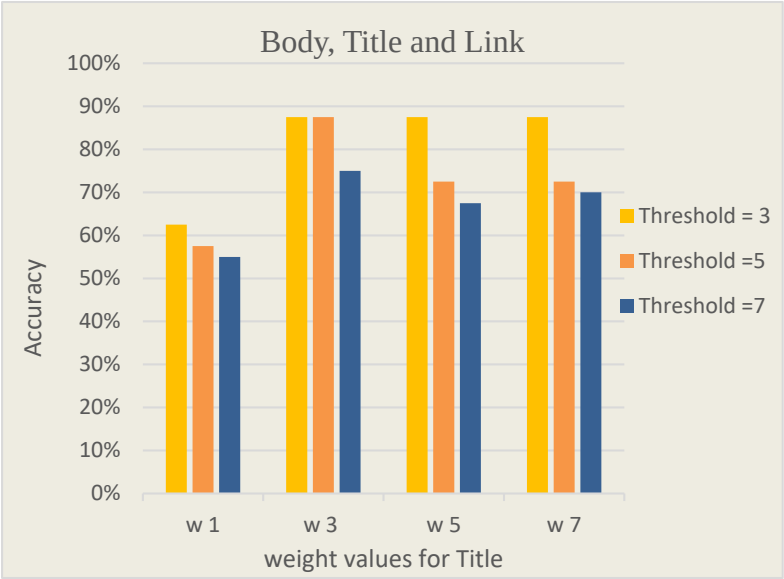

Fig. 2. Analysis results with various threshold values on BTL and Body using Enhanced C4.5 Classifier

X. CONCLUSION

The World Wide Web is the world's greatest collection of electronically accessible documents, making it the world's richest source of information. The issue with the Web is that the information is not clearly structured and arranged, making it difficult to find. Web page classification is important in this case for managing and retrieving relevant information from web resources. To increase the performance of the Web page categorization system, the presented Enhanced C4.5 algorithm considers the semantic class level and the

This publication is licensed under Creative Commons Attribution CC BY.

classification is based on not only content and but also title and link of the web pages. As a result, the proposed Semantic-based Web Page Categorization System can help to improve Web page classification performance.

ACKNOWLEDGMENT

I would like to greatly thank to Dr. Nyo Nyo Yee, Professor, Head of Department of Information Science, Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City), for her support and guidance. She gives me valuable advice for my research work. And I also would like to express heart fully thanks to Dr. Nandar Win Min, Associate Professor, Department of Information Science, Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City). She has provided helpful guidance during my research work.

REFERENCES

[1] Z. T. T. Myint and K. K. Win, "Triple Patterns Extraction from Unstructured Sentence Using Domain Specific Ontology", International Conference on Computer Applications (ICCA), 2013.

[2] H. H. Tar and T. T. S. Nyunt, "Ontology based Criterion in Categorical Clustering", International Conference on Computer Applications (ICCA), 2012.

[3] C. Goller, J. Loning and T. Will, "Automatic Document Classification", Munchen, Germany, pp. 145-161, 2000.

[4] https://en.wikipedia.org/wiki/Document_classification

[5] P. Phalak, K. Bhandari and R. Sharma, "Analysis of Decision Tree - A Survey", International Journal of Engineering Research & Technology, pp. 149-154, vol. 3, no. 3, March, 2014.

[6] B. R. Patel and K. K. Rana, "A Survey on Decision Tree Algorithm for Classification", International Journal of Engineering Development and Research (IJEDR), vol. 2, pp. 1-5, 2014.

[7] I. Bhuvana and C. Yamini, "Survey on Classification Algorithms for Data mining: (Comparison and Evaluation), International Journal of Advance Research in Science and Engineering, vol. 4, pp. 125-134, 2015.

[8] R. Revathy and R. Lawrance, "Comparative Analysis of C4.5 and C5.0 Algorithms on Crop Pest Data", International Journal of Innovative Research in Computer and Communication Engineering, vol. 5, pp. 50-58, 2017.

[9] B. Liu, "Web Data Mining", Department of Computer Science, University of Illinois at Chicago, USA, Springer-Verlag Berlin Heidelberg, 2007.

[10] Y. Chang and H. Huang, "Automatic Document Classifier System Based on Naïve Bayes Classifier and Ontology", Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, pp. 12-15, IEEE, 2008.

[11] W. K. Ong, J. L. Hong and F. Fauzi, "Ontological Based Web Page Classification", pp. 224-228, IEEE, 2012.

[12] M. Song, S. Lim and D. Kang, "Automatic Classification of Web Pages based on the Concept of Domain Ontology", Proceedings of the 12th Asia-Pacific Software Engineering conference, IEEE, 2005.

[13] https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a

AUTHORS

First Author – Hnin Pwint Myu Wai, University of Computer Studies (Bnamaw), Myanmar, hninpwintmyuwai14@gmail.com
Second Author – Dr. Nyo Nyo Yee, University of Technology (Yatanarpon Cyber City), Myanmar, nny1ster@gmail.com.
Third Author – Dr. Nandar Win Min, University of Technology (Yatanarpon Cyber City), Myanmar, nandarwinmin@gmail.com

This publication is licensed under Creative Commons Attribution CC BY.