Ontology based Web Page Classification System by using Enhanced C4.5 and Naïve Bayesian Classifiers

Hnin Pwint Myu Wai, Phyu Phyu Tar, Phyu Thwe
Faculty of Information and Communication Technology
University of Technology (Yatanarpon Cyber City)
Pyin Oo Lwin, Myanmar
hninpwintmyuwai14@gamil.com, thitagu7@gmail.com, pthwe19@gmail.com

***Abstract*—Today, web is a huge repository of information which needs for accurate automated classifiers for Web pages. Classification of Web page is essential to many tasks in Web information retrieval such as maintaining, web directories and focused crawling. So, this system proposes as the web page classification system based on semantic logic. For semantic, this system uses the ontology that stores each concept of each word. For classification, this system proposes the enhanced C4.5 decision tree and Naive Bayesian (NB) classifiers. In the original C4.5 classification algorithm, the traditional entropy measure is unable to measure the appropriateness of nodes when the class labels are the same. By using semantic technology, this system can effectively support to classify web pages into each category. To show the effectiveness, this system is tested by using HTML documents in the computer science domain.**

Keywords— semantic, classification, ontology, C4.5, NB

I. Introduction

Rapid development of technology leads human beings and the devices to connect to internet and share the data. Web is the largest collection of electronically accessible documents which make the richest source of information in the world. Thus, the information is accumulating in WWW at a very high rate. The problem with the Web is that this information is not well structured and organized so that it would be easily retrieved. In this scenario, it is necessary to categorize the web contents in an organized way. Automatic classification of web pages into relevant categories helps the search engines to give quicker and better results. Web page classification is the task of deciding whether a page belongs to a set of predefined category of document which is relevant to the topic.

With the drastic growth of web based information, web page classification process becomes one of the major challenges in organizing and maintaining the enormous collection of pages. Web page classification is useful for managing and extracting relevant information from Web content and in order to effectively use the knowledge available on the Web. This classification technique is employed to assign one or more predefined category labels to future Web pages.

To support the relevant information extraction, this system proposes the ontology based web page classification system by using enhanced C4.5 decision tree and Naive Bayesian classifiers. By using semantic technology, this system can effectively support to classify web pages into each category. Moreover, this system is motivated to overcome the weakness of the C4.5 decision tree classification algorithm. The proposed system is useful and essential for many tasks in web information retrieval such as maintaining web directories and focus crawling.

This paper is organized with six sections. Related work is described in section II. Classification is presented in section III. Proposed system design is presented in section IV. Experimental results are described in section V. Finally, conclusion is presented in section VI.

II. Related Work

In 2016, A. I. Saleh and M. F. A. Rahmawy [1] proposed a novel strategy for vertical web page classification, which is called classification using multilayered domain ontology (CMDO). It employed several web mining techniques, and depends mainly on proposed multi-layered domain ontology. In order to promote the classification accuracy, CMDO implied a distiller to reject pages related to other domains. CMDO also employed a novel classification technique, which is called graph based classification (GBC). The proposed GBC has pioneering features that other techniques do not have, such as outlier rejection and pruning. Experimental results have shown that CMDO outperforms recent techniques as it introduces better precision, recall, and classification accuracy.

In 2017, C. G. Fating and D. V. Zore [2] presented text classification that is the task of automatically sorting a set of documents into categories from a predefined set. Text classification is used to predict group membership for data instances within a given dataset. This paper introduced a new model based on probability and over all class frequency of term. Naive Bayesian classifier is based on Bayes theorem with independence assumptions between predictors. Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. The paper showed that the new probabilistic interpretation of tf × idf term weighting might lead to better understanding of statistical ranking mechanisms.

978-1-5386-7516-8/18/$31.00 ©2018 IEEE

In 2017, R. M. Rakholia and J. R. Saini [3] presented Naive Bayes (NB) statistical machine learning algorithm for classification of Gujarati documents. Six pre-defined categories sports, health, entertainment, business, astrology and spiritual are used for this work. A corpus of 280 Gujarat documents for each category is used for training and testing purpose of the categorizer. They have used k-fold cross validation to evaluate the performance of Naïve Bayes classifier. In this paper, experimental results show that the accuracy of NB classifier without and using features selection was 75.74% and 88.96% respectively.

III. Classification

Classification is a data mining technique used to predict group membership for data instances. There are many traditional classification methods like decision tree induction, k-nearest neighbor classifier, Bayesian networks, support vector machines, rule based classification, case-based reasoning, fuzzy logic techniques, genetic algorithm, rough set approach and so on. The basic difference between the algorithms depends on whether they are lazy learners or eager learners. The decision tree classifiers, Bayesian classifier, support vector classifier are eager learners as they use training tuples to construct the data model whereas nearest neighbor classifiers are lazy learners as they wait until a test tuple arrives for classification to perform generalization [4].

A. C4.5 Decision Tree Classifier

Decision trees are trees that classify instances by sorting them based on feature values [5]. In decision tree, each branch node represents a choice between number of alternatives and each leaf node represents a Decision. Decision trees are commonly used for gaining information for the purpose of Decision making. Decision tree starts with a root node which is for user to take actions. From this node, user split each node recursively according to Decision tree learning algorithms. Final result is a decision tree in which each branch represents a possible scenario of decision and its outcome. The widely used decision tree learning algorithms are ID3 (Iterative Dichotomiser), C4.5 and CART [6].

C4.5 is an extension of ID3 (Iterative Dichotomiser). This algorithm produces a decision tree for the given crop pest training data by recursively splitting that data. The decision trees generated by C4.5 can be used for classification since it is often referred to as a statistical classifier. C4.5 decision tree grows using Depth-first strategy. It allows pruning of the resulting decision trees. It can also deal with numeric attributes, missing values, and noisy data. In order to handle continuous attributes, it creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it. C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes [7].

1) Enhanced C4.5 Algorithm: For web page classification, this system proposes the enhanced C4.5 algorithm. This enhanced algorithm considers the semantic logic to choose the root node (best attribute). Based on the original and semantic classes, this algorithm calculates the normalized information gain for choosing best attribute. The enhanced C4.5 algorithm is as follows:

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 $a \in D$ 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. $a_{best}$ = Best attribute according to above computed criteria
11. Tree = Create a decision node that tests $a_{best}$ in the root
12. $D_v$ = Induced sub-datasets from $D$ based on a best
13. for all $D_v$ do
14. $Tree_v$ = C4.5($D_v$)
15. Attach $Tree_v$ to the corresponding branch of Tree
16. end for
17. return Tree
```

In the above enhanced C4.5 algorithm, step 6, 7 and 8 are different from the original C4.5 algorithm. The original C4.5 didn't consider the semantic class level. In the proposed enhanced C4.5 algorithm, both the original training class label and semantic class label are considered. Furthermore, this enhnaced algorithm calculates the gain ratio based on the training and semantic class label. So, the enhanced algorithm is more precise than the original algorithm.

1) Normalized Information Gain: This information gain is used to select the test attribute at each node in the tree. This method is as follows:

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

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

$$\text{Gain(A)} = \text{Info(D)} - \text{Info}_A\text{(D)} \qquad (3)$$

$$SplitInfo(A) = -\sum_{i=1}^{m} \frac{|Ci|}{C} \log 2 \frac{|Ci|}{C} \qquad (4)$$

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

In the normalized information gain, 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. It is also known as Entropy of 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. Gain(A) and Gain ratio (A) shows how much would be gained by branching on A. The attribute A with the highest information gain, Gain ratio (A), is chosen as the splitting attribute at Node N [7].

2) Advantages of C4.5 Algorithm: The advantages of C4.5 algorithm are as follows:

- C4.5 algorithms construct trees and grow its branches.
- The attribute with the highest normalized information gain is chosen to make the decision.
- This algorithm is used to handle continuous and discrete values [9].

B. Naive Bayesian (NB) Classifier

Naive Bayes classifier is simple classifier which is based on Bayes Theorem of conditional probability and strong independence assumptions. This classifier emphasizes on measure of probability that whether the document A belongs to class B or not. It is non sensitive to irrelevant features. It is used in personal email sorting, document categorization, email spam detection and sentiment detection [10].

1) Processing Steps of NB Classifier: The processing steps of Naïve Bayesian classifier are as follows:

1. Each data sample is represented by n-dimensional feature vector, $X=(x_1, x_2\dots x_n)$ depicting n-measurements made on the sample from n- attributes, respectively, $A_1, A_2,\dots., A_{n.}$
2. Suppose that there are m classes, $C_1, C_{2,}\dots, C_m$. Given an unknown data sample X, Naïve Bayesian classifier assigns an unknown sample X to the class C_i if and only if

$$P(C_i\backslash X) > P(C_j\backslash X) \text{ for } 1 \le j \le m, j \ne i \quad (6)$$

The class C_i for which $P(C_i\backslash X)$ that is maximized, called maximum posteriori hypothesis. By Bayes theorem,

$$P(C_i\backslash X) = P(X\backslash C_i)\, P(C_i)/\, P(X) \quad (7)$$

3. As P(X) is constant for all classes, only $P(X\backslash C_i)\, P(C_i)$ need to be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) =\dots\dots = P(C_m)$ and we would therefore maximize $P(X\backslash C_i)\, P(C_i)$.
4. Given data sets with many attributes, it would be extremely expensive to compute $P(X\backslash C_i)$. In order to reduce computation in evaluating $P(X\backslash C_i)$, the Naïve assumption of class conditional independence is made. Thus,

$$P(X\backslash C_i) = \prod\nolimits^{n}_{k=1} P(x_k\backslash C_i) \quad (8)$$

The probability $P(x_1\backslash C_i), P(x_2\backslash C_i),\dots P(x_n\backslash C_i)$ can be estimated from the data samples.

5. In order to classify an unknown sample X, $P(X\backslash C_i)\, P(C_i)$ is evaluated for each class $C_{i,.}$ Sample X is then assigned to the class C_i if and only if

$$P(X\backslash C_i)\, P(C_i) > P(X\backslash C_j)\, P(C_j) \quad \text{for } 1 \le j \le m, j \ne i \quad (9)$$

In other words, it is assigned to the class C_i for which $P(X\backslash C_i)\, P(C_i)$ is the maximum [8].

2) Advantages of NB Classifier: The advantages of NB classifier are as follows:

- Training is very easy and fast.
- NB Classifier only requires a small amount of training data to estimate the parameters necessary for classification.
- NB classifiers have worked quite well in many complex real world problems.
- NB algorithm affords fast highly scalable model building and scoring.
- NB classifier can be used for both binary and multiclass classification problems.

C. Ontology

Ontology is an explicit formal conceptualization of some domain of interest. Ontology renders shared vocabulary and taxonomy which models a domain with the definition of objects and/or concepts and their properties and relations. Using ontology as a controlled vocabulary, accuracy value can be improved in retrieving semantic information. Ontology is an information model containing concepts and relations. Ontology is increasingly used in various fields such as knowledge management, information extraction, and the semantic web [11].

IV. System Desigh

Proposed system design about ontology based web page classification system is shown in Fig.1.

In the proposed system, there are three main parts. These are pre-processing, semantic extraction and classification. In the pre-processing part, this system performs the tokenization and stopwords removal process to the user inputted web page and training web pages. After performing the stopwords removal process, this system extracts the features from these web pages. In the semantic extraction part, this system searches the semantic meaning that is relevant to the extracted features. For semantic extraction, this system uses the ontology.

Then, this system performs the classification process by using the enhanced C4.5 algorithm. This enhanced algorithm performs the classification process based on both the original class and the extracted semantic class. According to this algorithm, this system classifies the training web pages to produce the decision tree. By using the decision tree, this system produces the classification decision rules. Sometimes, this system can face the problem in the decision tree production process. This problem is that both the enhanced C4.5 and original C4.5 algorithms can't sometimes produce the last leaf nodes as the class label. In this situation, this system solves this problem by using Naive Bayesian classifier. So, this system uses both the enhanced C4.5 and Naïve Bayesian classifiers. Finally, this system assigns the category about the user inputted web page according to the decision rules.

A. Explanation of the System

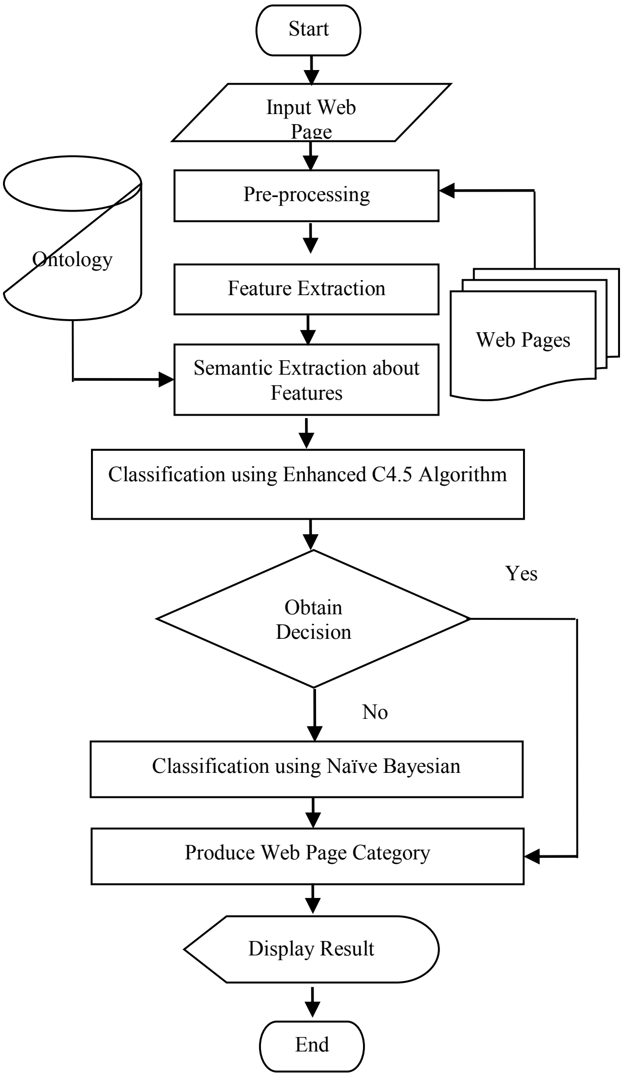

Start
Input Web Page
Ontology
Pre-processing
Feature Extraction
Web Pages
Semantic Extraction about Features
Classification using Enhanced C4.5 Algorithm
Obtain Decision
Yes
No
Classification using Naïve Bayesian
Produce Web Page Category
Display Result
End

Fig.1. Proposed System Design

In the explanation of the system, the sample five web pages are used with three categories such as "web data mining", "cryptography" and "distributed system". These sample web pages are as follows:

- Web page 1: PageRank has emerged as the dominant link analysis model for Web search, partly due to its query independent evaluation of Web pages.
- Web page 2: HITS is search query dependent to retrieve information. When the user issues a search query, HITS first expands the list of relevant pages that include information.
- Web page 3: Web information retrieval is the study of helping user to find information that matches their query. On the web, the transmitted information that is relevant query needs to be processed into an unrecognizable form in order to be secured.
- Web page 4: With the fast progression of information exchange in electronic way, security is becoming more important in information transmission as well as in storage. RC4 algorithm protects the confidential data from unauthorized access. RC4 is the encryption method.
- Web page 5: Web is a collection of host machines and server, which delivers information that is relevant query. On web, query related information are distributed to client.

By using these training five web pages, this system firstly performs the tokenization and stopwords removal process. Then, this system extracts the keyword features according to the threshold value. In this sample, the threshold value is 2. For the keyword features, this system extracts the relevant semantic features by using ontology.

Then, this system performs the classification process by using both keyword and semantic features. For this classification process, this system also considers the original class and semantic class. Training data for classification is shown in Table 1.

For classification, this system uses the enhanced C4.5 algorithm. This classification algorithm firstly creates the decision tree to produce the decision rules. For decision tree, this system performs each iteration by choosing the attribute that has highest gain ratio result as the root node.

The "RC4" is the root node in the first iteration because it has highest gain ratio result. In the second iteration, root node is "Information". In the third iteration, root node is "HITS". After calculating gain ratio for each iteration, this system finally produces the decision tree. In this decision tree, this system faces the problem for producing decision rules. From the node "HITS", this system can't choose the class level because this system faces two class levels that are "cryptography" and "distributed system".

TABLE I. TRAINING DATA FOR CLASSIFICATION

Web Page	web	information	query	HITS	RC4	encryption	Class
1	Yes	No	No	No	No	No	web data mining
2	No	Yes	Yes	Yes	No	No	web data mining
3	Yes	Yes	Yes	No	No	No	cryptograph y
4	No	Yes	No	No	Yes	Yes	cryptograph y
5	Yes	Yes	Yes	No	No	No	distributed system

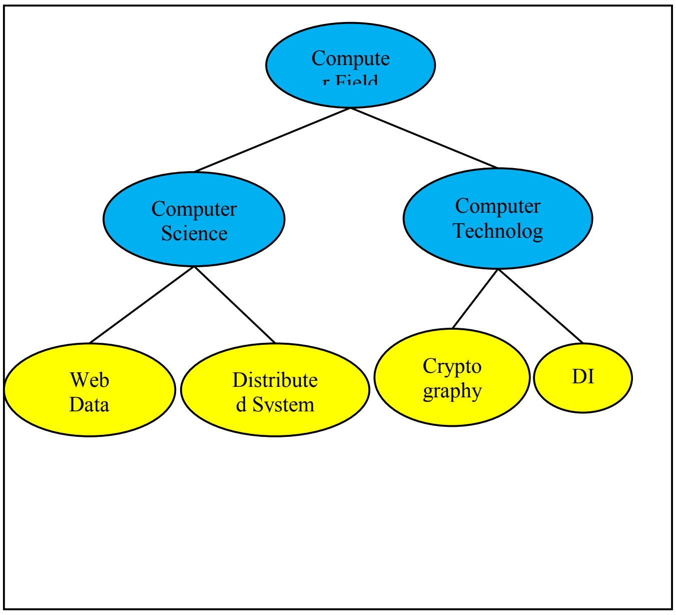
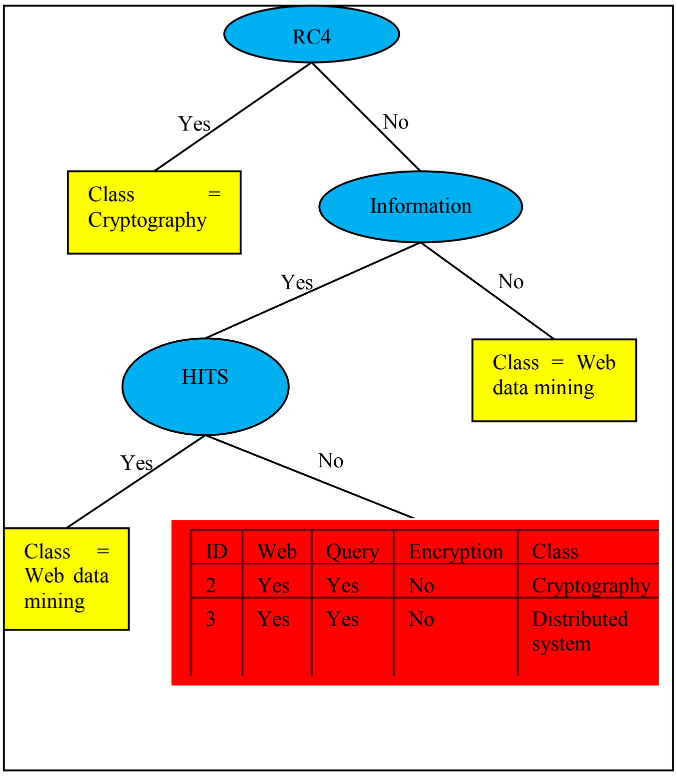


Fig.2. Semantic and Original Class Levels



ID	Web	Query	Encryption	Class
2	Yes	Yes	No	Cryptography
3	Yes	Yes	No	Distributed system

Fig.3. Decision Tree using Enhanced C4.5 Algorithm

In this situation, this system solves this problem by using Naive Bayes classifier. After solving, this system produces the decision rules for classification. Fig. 3 shows the decision tree using the enhanced C4.5 algorithm. Fig. 4 shows the decision tree using the enhanced C4.5 and Naive Bayesian classifier.

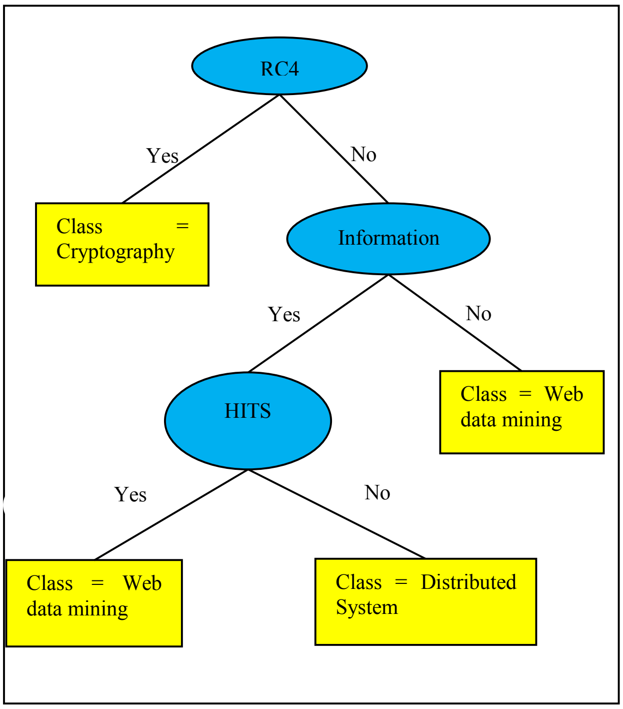


Fig.4. Decision Tree using Enhanced C4.5 and Naïve Bayesian Classifiers

After obtaining decision tree, this system produces the decision rules for classification. Decision rules are as follows:

- Rule 1: **IF** “RC4 = yes”
  THEN Class = “Cryptography”
- Rule 2: **IF** “RC4 = no”
  AND “information = no”
  THEN Class = “web data mining”
- Rule 3: **IF** “RC4 = no”
  AND “information = yes”
  AND “HITS = yes”
  THEN Class = “web data mining”
- Rule 4: **IF** “RC4 = no”
  AND “information = yes”
  AND “HITS = no”
  THEN Class = “distributed system”.

Using decision rules, this system classifies the future web pages.

V. Experimental Results

In this system, holdout method is used to measure the performance of the system. In the holdout method, the given data are randomly partitioned into two independent sets, a training set and a test set. Typically, two thirds of the data are allocated to the training set, and the remaining one third is allocated to the test set. The training set is used to derive the classifier, whose accuracy is estimated with the test set. The classification accuracy A_i of classifier is evaluated by the formula

$$A_i = t/n * 100 \tag{10}$$

where A_i is accuracy of the classifier, t is the number of testing data correctly classified and n is the total number of testing data.

This system is tested by using 150 web pages and 200 web pages to measure the accuracy of the proposed ontology based web page classification system. For testing, this system used the decision rules from the web page classification. From the 150 web pages testing, this system occurs the correct rate is 92% and the error rate is 8%. And then, the correct rate and error rate of 200 web pages is 92.5% and 7.5% respectively. Experimental result of the system is shown in Fig. 5.

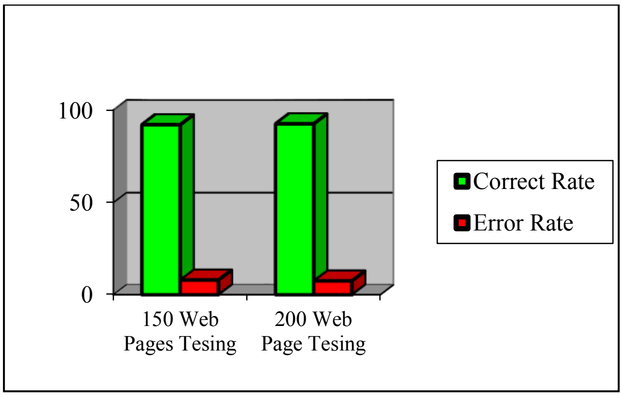


Fig.5. Experimental Results of the System

VI. Conclusion

The proposed ontology based web page classification system is not only used to improve the performance of search engines but also essential for the development of web directories. To assign one or more predefined category labels for future web pages, this system used the enhanced C4.5 algorithm by considering the semantic logic. Moreover, this system can solve the problem that is faced by the decision tree algorithms. For problem solving, this system used the Naïve Bayesian classifier. Therefore, the proposed ontology based web page classification system can support to enhance the performance of web search engine.

Acknowledgment

Firstly, the author would like to express her profound gratitude to Dr. Aung Win, Rector of University of Technology (Yatanarpon Cyber City), for his invaluable directions and managements. The author would like to acknowledge the effective support of her honorable supervisor Dr. Phyu Phyu Tar, and Dr. Phyu Thwe as a co- Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City), who have supported closely with her at every stage in her research work and valuable suggestions, true line guidance, supervision and editing this research.

References

[1] A. I. Saleh and M. F. A. Rahmawy, "A Semantic based Web Page Classification Strategy using Multi-Layered Domain Ontology", Springer, Science and Business Media, New York, 2016.

[2] R. M. Rakholia and J. R. Saini, "Classification of Gujarati Documents using Naïve Bayes Classifier", Indian Journal of Science and Technology, vol. 5, pp. 1-9, 2017.

[3] M. S. Vani, A. Sherin and K. Saranya, "Survey on Classifcation Techniques used in Data Mining and their Recent Advancements", International Journal of Science, Engineering and Technology Research, vol. 3, no. 9, 2014.

[4] N. P. Thair, "Survey of Classification Techniques in Data Mining", Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS), vol. 1, 2009.

[5] S. N. Chary and B. Rama, "A Survey on Comparative Analysis of Decision Tree Algorithms in Data Mining", International Conference on Innovative Applications in Engineering and Information Technology (ICIAEIT), vol. 3, no. 1, pp. 91-95, 2017.

[6] 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, no. , vol. 5, pp. 50-58, 2017.

[7] H. Jiawei and K. Micheline, “Data Mining Concepts and Techniques”, Simon Fraser University, USA, 2001.

[8] D. Sindhuja and R. J. Priyadarsini, "A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder", vol. 5, no. 5, pg. 483-488, 2016.

[9] R. P. Rajeswari and K. Juliet, "Text Classification for Student Data Set Using Naive Bayes Classifier and KNN Classifier", International Journal of Computer Trends and Technology (IJCTT), vol. 43, no. 1, pp. 8-12, 2017.

[10] J. Brank, M. Grobelnik and D. Mladenic, “A Survey of Ontology Evaluation Techniques”, Department of Knowledge Technologies Jozef Stefan Institute, Slovenia, 2008.

[11] J. Brank, M. Grobelink and D.Mladenic, “A Survey of Ontology Evaluation Techniques”, Department of Knowledge Technologies, Jozef Stefan Institute, Slovenia, 2008.