Performance Comparison between Keyword-based and Semantic-based Web Page Classification Systems

Hnin Pwint Myu Wai[*], Phyu Phyu Tar[*], Phyu Thwe[**]

Faculty of Information & Communication Technology
University of Technology (Yatanarpon Cyber City)
hninpwintmyuwai14@gmail.com, thitagu7@gamil.com, pthwe19@gmail.com

DOI: 10.29322/IJSRP.9.03.2019.p8751
http://dx.doi.org/10.29322/IJSRP.9.03.2019.p8751

Abstract- With the increase of information, web page classification as one of the methods of text mining plays vital role in many management and organizing information. Web page classification is the process of assigning a web page to one or more predefined category labels. With the high availability of information from diverse sources, web page classification tasks have attained paramount importance. The traditional web page classification usually ignores the semantic relations among the keywords or web pages. To effectively solve the semantic problem, the ontology is used to capture the semantic information among web pages. So, this system proposes the semantic-based web page classification system by using NB-based C4.5 algorithm. This algorithm can solve the barrier from both the C4.5 algorithm and keyword-based web page classification. To know which classifier is more than other, this system compares the performance between keyword-based and semantic-based web page classification.

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

I. Introduction

With the increasing availability of digital documents from diverse sources, web page classification is gaining popularity day in and day out. There is a mushroom growth of digital data made available in the last few years, data discovery and data mining have worked together to extract meaningful data into useful information and knowledge. Text mining refers to the process of deriving high quality information from text. It is conducive in utilizing information contained in web pages in various ways including discovery of patterns, association among entities etc. and this is done with the amalgamation of Natural Language Processing, Data Mining and Machine learning techniques.

Infeasibility of human beings to go through all the available web pages to find the web page of interest precipitated the rise of web page classification. Automatically classifying web pages could provide people a significant ease in this realm. Text classification assigns web pages into one or more predefined categories. The notion of classification is very general and has many applications within and beyond information retrieval (IR). For instance, text classification finds its application in automatic spam detection, sentiment analysis, automatic detection of obscenity, personal email sorting and Topic specific or Vertical Searches.

Several techniques included in text classification are: Naive Bayes classifier, Expectation mining, SVM, Artificial Neural Network, concept mining, decision trees, etc. These techniques have their individual set of advantages which make them suitable in almost all classification jobs. Among them, this system uses the decision tree classifier that is C4.5 algorithm. To solve the semantic problem, this system constructs the new supervised classifier for web page classification. This classifier is NB (Naïve Bayesian)-based C4.5 algorithm. This algorithm can solve the problem about the original C4.5 classifier and keyword-based web page classification.

This paper is organized with nine sections. Related work is described in section II. Text classification is presented in section III. Keyword-based and semantic-based web page classifications are described in section IV and V. Proposed system design is presented in section VI. Explanation and experimental results are described in section VII and VIII. Finally, conclusion is presented in section IX.

II. RELATED WORK

In 2006, H. Zhang and H. Song [8] presented fuzzy related classification approach based on semantic measurement for web document. In this paper, a novel approach of automatically classification for web document based on ontology concept semantic measurement is proposed, in which a method to compute the semantic similarity between concepts is put forward, and the fuzzy related technology combined with vector space model in classifying Web documents into a predefined set of ontology categories is adopted.

In 2006, S. Deng and H. Peng [9] described document classification based on support vector machine using a concept vector model. The traditional document classification usually ignores the semantic relations among the keywords or documents. To effectively solve the semantic problem, the domain ontology is used to capture the semantic information among different terms or keywords in the documents. Using the concept vector model (CVM), domain-related semantic information more exactly from documents can be extracted. In the model, concept vector is extracted from a document by the matching method. According to concept features of the documents, documents are classified into a suitable category by support vector machine (SVM). The experimental results show that their CVM method yields higher accuracy compared to the traditional term-based vector space model (VSM) methods.

In 2010, R.Mohamed and J. Watada [10] presented an evidential reasoning based latent semantic analysis (LSA) approach to document classification for knowledge acquisition. Web is one of major information resources. Failure in proper management of knowledge leads to incorrect result returned by search engines. Therefore, the web should have an effective information retrieval system to improve the correctness of retrieval results. This study provides a method to assign a new document to the fittest category out of predefined categories, where latent semantic analysis (LSA) is used to evaluate each term in documents, the similarity between terms and documents as well as the one between terms and categories. The objective of their method is to fuse evidential reasoning method with LSA which can assign a new document to a predefined category. The method provides better results in performance of classification comparing to the fusion of an evidential reasoning approach with term frequency inverse document frequency (TFIDF).

III. TEXT CLASSIFICATION

Text classification is used to classify the web page of similar types. Text classification can be also performed under supervision i.e. it is a supervised leaning technique. Text classification is a process in which web pages are sorted spontaneously into different classes using predefined set. It refers to resolving the issue to identify web pages based on their content into a definite number of predefined classes. Text classification plays a significant role in wide variety of fields such as information retrieval, web pages classification and so on. Text classification process includes the following steps [1]:

- Document Collection: The initial step is to collect different formats of documents such as html, web content, .pdf, .doc, etc.
- Pre-processing: It consists of two steps. First step is tokenization that presents text document into word format. In second step, text is represented by the number of features to conflate the token to their root format, e.g. computing to compute.
- Indexing: It is used to minimize the complexity of document and model them easy to handle.
- Classification: Automatic classification of documents in predefined categories is gaining active attention of many researchers. Supervised, unsupervised and semi-supervised methods are used to classify documents [2].

There are many 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 [3].

IV. KEYWORD-BASED WEB PAGE CLASSIFICATION

For the keyword-based web page classification, this system uses the decision tree classifier. There are many decision tree classifiers that are iterative dichotomiser (ID3), extended version of ID3 (C4.5), classification and regression tree (CART), chi-squared automatic interaction detector (CHAID) and multivariate adaptive regression splines (MARS). Among them, this system uses the C4.5 classification algorithm. The C4.5 classifier generates a decision tree for the given data by recursively splitting that data. The decision tree grows using depth-first strategy [5]. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub-lists [4].

The C4.5 classification algorithm is as follows:

Algorithm : C4.5 Decision Tree

Input: An attribute-valued dataset *D*

```
Tree = { }
if $D$ is "pure" OR other stopping criteria met then
    Terminate
end if
for all attribute $a$ € $D$ do
    Compute gain ratio if we split on $a$ based on class level
end for
$a_{best}$ = Best attribute according to above computed criteria
Tree = Create a decision node that tests $a_{best}$ in the root
$D_v$ = Induced sub-datasets from $D$ based on a best
for all $D_v$ do
    $\text{Tree}_v$ = C4.5($D_v$)
    Attach $\text{Tree}_v$ to the corresponding branch of Tree
end for
return Tree
```

A. *Normalized Information Gain*

The normalized information gain is used to select the test attribute at each node in the tree. This method is as follows:

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D) \tag{1}$$

$$\mathrm{Info}(D) = -\sum_{i=1}^{m} P_i \mathrm{Log}_2 (P_i) \qquad (2)$$

$$\mathrm{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \mathrm{Info}(D_j) \qquad (3)$$

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

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

Pi 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. |Dj |/|D | acts as the weight of the jth partition. InfoA(D) is the expected information required to classify a tuple from D based on the partitioning by A. Ci is the objects in class C that have value A of Ai. 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 [6].

V. SEMANTIC-BASED WEB PAGE CLASSIFICATION

For semantic-based web page classification, this system proposes the Naïve Bayesian (NB) based C4.5 algorithm. To increase the performance of web page classification system, this algorithm produces the decision tree by classifying each web page with the semantic and original class label. For semantic logic, this system creates the ontology that includes semantic classes. The NB-based C4.5 algorithm consists of two parts. The first part is the decision tree based classification. Sometimes, the decision tree can face the problem to produce the last leaf node as the class label. In this situation, this system solves this problem according to the probability based classification. This probability based classification is the second part of NB-based C4.5 algorithm. Finally, this system assigns the category about the user inputted web page according to the decision rules.

The NB-based C4.5 algorithm is as follows:

Algorithm : NB-based C4.5 Classifier

Input: An attribute-valued dataset *D*

First Part: Decision Tree-based Classification

```
Tree = { }
if D is "pure" OR other stopping criteria met then
    Terminate
end if
for all attribute a € D do
    Compute gain ratio if we split on a based on original class level
    Next, Compute gain ratio about a based on semantic class level
    Combine each gain ratio result that is obtained by calculating each attribute value based on original class level and semantic class level
end for
$a_{best}$ = Best attribute according to above computed criteria
Tree = Create a decision node that tests $a_{best}$ in the root
$D_v$ = Induced sub-datasets from D based on a best
for all $D_v$ do
    $Tree_v$ = C4.5($D_v$)
    Attach $Tree_v$ to the corresponding branch of Tree
end for
return Tree
```

Second Part: Probability-based Classification

1. Each data sample is represented by n-dimensional feature vector, X=(x_1, x_2… x_n) depicting n-measurements made on the sample from n- attributes, A_1, A_2,.., A_n.

2. Suppose that there are m classes, C1, C2, …, Cm. Given an unknown data sample X, classifier assigns an unknown sample X to the class Ci if and only if

$$P(Ci \backslash X) > P(Cj \backslash X) \text{ for } 1 \leq j \leq m, j \neq i \qquad (6)$$

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

$$P(Ci \backslash X) = P(X \backslash Ci)\, P(Ci) / P(X) \qquad (7)$$

4. As P(X) is constant for all classes, only P(X\Ci) P(Ci) 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 (C1) = P (C2) =…. = P(Cm) and we would therefore maximize P(X\Ci) P(Ci).

5. Given data sets with many attributes, it would be extremely expensive to compute P (X\Ci). In order to reduce computation in evaluating P(X\Ci), the Naïve assumption of class conditional independence is made.

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{8}$$

The probability P(x1\Ci), P(x2\Ci),… P(xn\Ci) can be estimated from the data samples.

6. In order to classify an unknown sample X, P (X\Ci) P (Ci) is evaluated for each class Ci,. Sample X is then assigned to the class Ci if and only if

$$P(X\backslash Ci)\ P(Ci) > P(X\backslash Cj)\ P(Cj) \quad \text{for } 1 \leq j \leq m, j \neq i \tag{9}$$

In other words, it is assigned to the class Ci for which P(X\Ci) P(Ci) is the maximum.

A. Ontology

Ontology can be viewed as a declarative model of a domain that defines and represents the concepts existing in that domain, their attributes and relationships between them. It is typically represented as a knowledge base which then becomes available to applications that need to use and/or share the knowledge of a domain. Ontology has been adopted in many business and scientific communities as a way to share, reuse and process domain knowledge. Ontology plays a major role in the development of the semantic Web.

VI. PROPOSED SYSTEM DESIGN

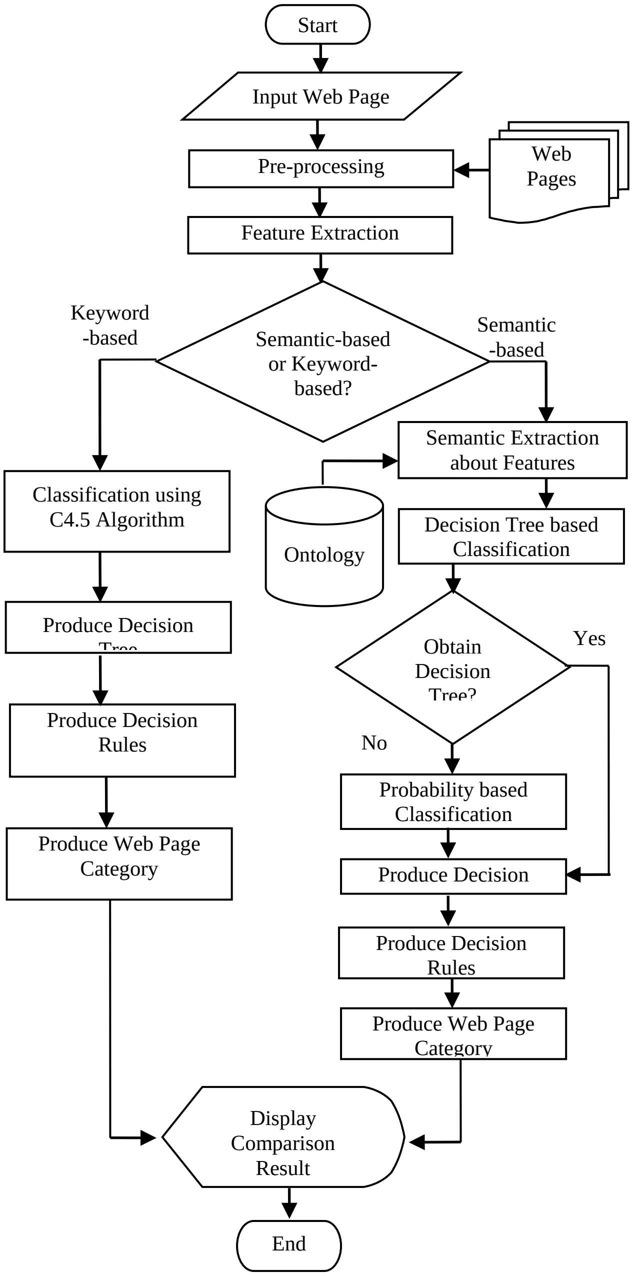

Start
Input Web Page
Pre-processing
Web Pages
Feature Extraction
Keyword -based
Semantic-based or Keyword-based?
Semantic -based
Semantic Extraction about Features
Classification using C4.5 Algorithm
Ontology
Decision Tree based Classification
Produce Decision Tree
Obtain Decision Tree?
Yes
No
Produce Decision Rules
Probability based Classification
Produce Web Page Category
Produce Decision
Produce Decision Rules
Produce Web Page Category
Display Comparison Result
End

Figure 1: System Design

The proposed system compares the performance between the keyword-based and semantic-based web page classification. At first, this system performs the pre-processing that includes the tokenization, stopwords removal by using input web page and training web pages. Then, this system extracts the features from each web page. In this system, the user can use not only the semantic-based web page classification process but also the keyword-based web page classification process. Semantic-based process includes five sub-processes that are semantic feature extraction, decision tree classification, probability based classification, decision tree production and decision rules production. But, the keyword-based process includes three sub-processes that are C4.5 classification, decision tree production and decision rules production. Using decision rules, these two processes classify the input web page into the relevant category (class). Finally, this system compares the accuracy results to know which classifier is better between these two classification processes. Proposed system design is shown in Figure 1.

VII. EXPLANATION OF THE SYSTEM

As a sample, this system is tested by using five web pages with three categories (web data mining, cryptography and distributed system). These 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.

For classification, this system firstly performs the tokenization and stopwords removal for each training web page. Then, this system extracts the keyword features. To choose these features, this system defines threshold value. In this sample, threshold value is 2. From each web page, this system extracts each keyword feature.

Table I: Training Data

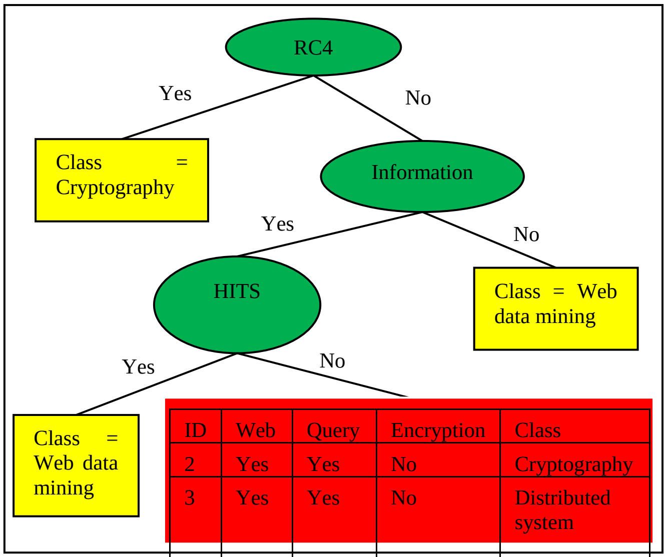
Web Page	web	information	query	HITS	RC4	encryption	Class
1	Yes	No	No	No	No	No	web data mining
2	No	Yes	Yes	Yes	No	No	web data mining
3	Yes	Yes	Yes	No	No	No	cryptography
4	No	Yes	No	No	Yes	Yes	cryptography
5	Yes	Yes	Yes	No	No	No	distributed system

Training data is shown in Table I. These training data are collected according to the feature extraction process.

A. *Keyword-based Classification Process*

Using these extracted keyword features, this system classifies training web pages, and produces the decision trees and rules by using C4.5 algorithm. To create the decision tree, this system calculates the gain ratio for each attribute. Attribute that has highest gain ratio result is chosen as root node. After calculating gain ratio for each iteration, this system finally produces the decision tree. Using decision tree, this system produces the decision rules to classify the future web pages.

In the C4.5 decision tree production, this system faces the problem about the leaf node production. Due to the some training data, this keyword-based web page classification cannot produce the complete decision tree. So, decision rules aren't completely corrected. These results are shown in Figure 2.



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

Figure 2: Decision Tree using C4.5 Classifier

B. *Semantic-based Classification Process*

For the semantic-based web page classification, this system uses the NB-based C4.5 classifier. This classifier produces the decision tree and rules by using both the keyword class and the semantic class. For semantic, this system uses the ontology. Moreover, this NB-based C4.5 classifier can solve the limitation of the original C4.5 classifier.

Table II: Training Data from the Semantic View

Web Page	web	information	query	HITS	RC4	encryption	Class
1	Yes	No	No	No	No	No	Computer science
2	No	Yes	Yes	Yes	No	No	Computer science
3	Yes	Yes	Yes	No	No	No	Computer technology
4	No	Yes	No	No	Yes	Yes	Computer technology
5	Yes	Yes	Yes	No	No	No	Computer science

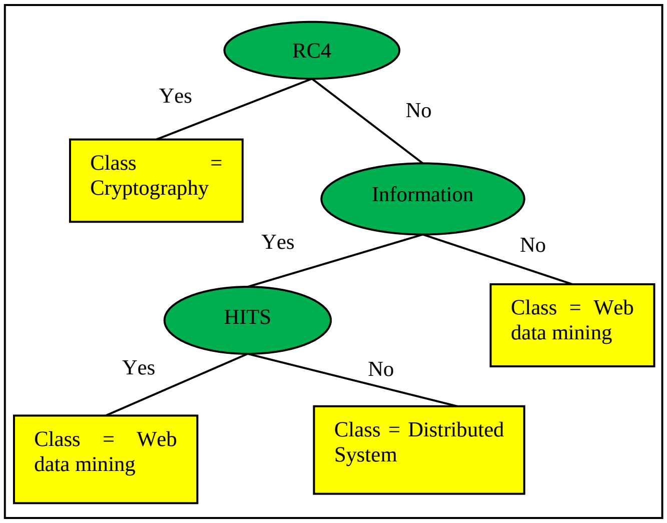


Figure 3: Decision Tree using NB-based C4.5 Classifier

Training data from the semantic view is shown in Table 2. Decision tree using NB-based C4.5 classifier is shown in Figure 3.

C. *Decision Rules*

According to the 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. In the decision rules production process, the keyword-based classifier cannot produce the rule 4. But, the semantic-based classifier can solve this problem by producing the classification rules.

VIII. EXPERIMENTAL RESULTS OF THE SYSTEM

To measure the performance of this system, this system uses the holdout method. It is the simplest kind of cross validation. The data set is separated into two sets, called the training set and the testing set. The function approximator fits a function using the training set only. Then the function approximator is asked to predict the output values for the data in the testing set. The errors it makes are accumulated as before to give the mean absolute test set error, which is used to evaluate the model. The advantage of this method is that it is usually preferable to the residual method and takes no longer to compute. However, its evaluation can have a high variance. The evaluation may depend heavily on which data points end up in the training set and which end up in the test set.

The classification accuracy Ai of classifier is evaluated by the formula

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

where Ai is accuracy of the classifier, t is the number of testing data correctly classified and n is the total number of testing data.
To measure the performance of keyword-based and semantic-based web page classification, this system is tested by using 200 web pages. For measurement, this system uses the decision rules that are obtained from the classification about 200 web pages. The experimental results of the keyword-based and semantic-based web page classification are shown in Table III and IV. The performance comparison results about correct rate and error rate are shown in Figure 4 and 5.

Table III: Experimental Results of the Keyword-based Web Page Classification

Test ID	*No of Web Pages*		*Accuracy (%)*	
	Training Web Pages	*Testing Web Pages*	*Correct Rate*	*Error Rate*
Test 1	120	90	85	15
Test 2	110	80	84	16
Test 3	130	70	86	14
Test 4	135	65	89	11

Table IV: Experimental Results of the Semantic-based Web Page Classification

Test ID	*No of Web Pages*		*Accuracy (%)*	
	Training Web Pages	*Testing Web Pages*	*Correct Rate*	*Error Rate*
Test 1	120	90	95	5
Test 2	110	80	94	6
Test 3	130	70	97	3
Test 4	135	65	97	3

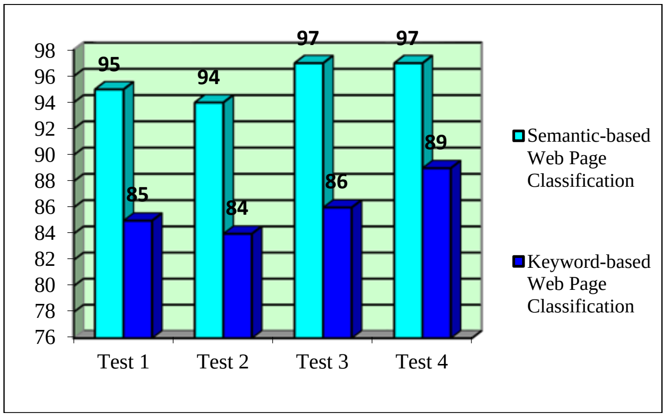


Figure 4: Performance Comparison Result about Correct Rate

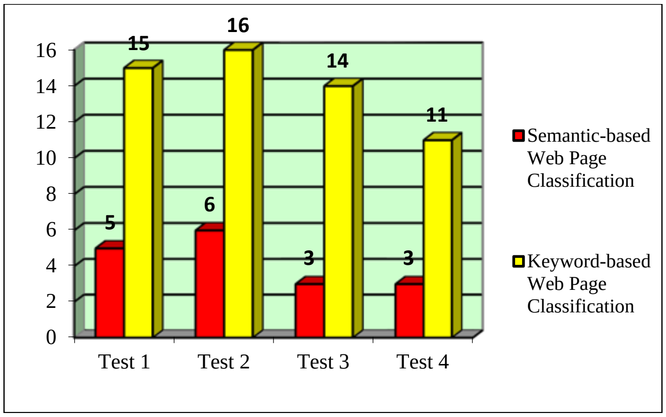


Figure 5: Performance Comparison Result about Error Rate

IX. CONCLUSION

The proposed system supports to enhance the performance of web search engine. This web page classification system is essential for the development of web directories. To assign one or more predefined category labels for future web pages, this system proposed the NB (Naïve Bayesian)-based C4.5 algorithm by considering the semantic logic. The proposed semantic-based classifier can point out the lack of keyword-based classifier and original C4.5 classifier. This original classifier sometimes faces the problem about decision rules production. So, this system compared and pointed out the performance between the semantic-based and keyword-based web page classification.

REFERENCES

[1] B. Kaur and G. Bathla, "Document Classification using Various Classification Algorithms: A Survey", International Journal on Future Revolution in Computer Science & Communication Enineering (IJFRCSCE), pp. 150-155, vol. 4, no. 2, 2018.

[2] U. Singh and S. Hasan, "Survey Paper on Document Classification and Classifiers", International Journal of Computer Science Trends and Technology (IJCST), vol. 3, no. 2, pp. 83-87, 2015.

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

[4] Z. Xiaoliang and W. Jian, "Research and Application of the improved Algorithm C4.5 on Decision Tree", International Conference on Test and Measurement, pp. 184-187, IEEE, 2009.

[5] S. Singh and P. Gupta, "Compartive Study ID3, CART and C4.5 Decision Tree Algorithm: A Survey", International Journal of Advanced Information Science and Technology (IJAIST), pp. 97-103, vol. 27, no. 27, 2014.

[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] Y. C. Kiong, S. Palaniappan and N. A. Yahaya, "Health Ontology System", 7th International Conference on IT in Asia (CITA), IEEE, 2011.

[8] H. Zhang and H. Song, "Fuzzy Related Classification Approach based on Semantic Measurement for Web Document", 6th International Conference on Data Mining - Workshops (ICDMW), IEEE, 2006.

[9] S. Deng and H. Peng, "Document Classification based on Support Vector Machine using A Concept Vector Model", Proceedings of the International Conference on Web Intelligence, IEEE, 2006.

[10] R.Mohamed and J. Watada, "An Evidential Reasoning based LSA Approach to Document Classification for Know Acquisition", Proceedings of the 2010 IEEE IEEM, pp. 1092-1096, IEEE, 2010.