# Multi-category Classification of Web Pages by using Random Forest Classifier

Win Thanda Aung
*University of Computer Studies, Yangon*
*winthanda.ucsy@gmail.com*

Thi Thi Soe Nyunt
*University of Computer Studies, Yangon*
*ttsoenyunt@gmail.com*

## Abstract

*To classify Web objects into predefined semantic structure is called the Web Page classification. One of the most essential technique for Web Mining is the automatic web page classification given that the web is a huge repository of various information including images, videos etc. And there is a need for categorization web pages to satisfy user needs. The classification of web pages into each category exclusively relies on man power which cost much time and effort. To alleviate this manually classification problem, more researchers focus on the issue of web pages classification technology. In this paper, we proposed Random Forest Classifier (RF) based on random forest method for multi-category web page classification. The proposed RF classifier can classify web pages efficiently according to their corresponding class without using other feature selection methods. We compared the accuracy of the proposed approach to decision tree classifier using in the same Yahoo web pages. The experiments have shown that the proposed approach is suitable for the multi-category web page classification.*

## 1. Introduction

To classify the information on Internet into a certain number of pre-defined categories is the goal of web page classification. Nowadays, there are so many thousands of web pages are increasing significantly on the World Wide Web. As a result of this, there are very important to mine the specific information on the Web. The task to find Web pages which present information satisfying user requirement is difficult. Classification is an essential tool that allows a person visiting a website to navigate it quickly and efficiently. Web pages classification can also help improve the quality of web search.

There are many different dimensionality and different representing forms of heterogeneous data sources in Web pages. In order to classify this data source into suitable classes, researchers proposed many classification algorithms. Neural network method is also applicable to multivariate non-linear problems. The transformations of the variables are automated in the computational process. This method is minimizing over fitting requires a great deal of computational effort. But neural network is that they are notoriously slow when so many training data sets exist. Random Forest method can handle hundreds or thousands of input data and faster than other methods.

One of the supervised method is a support vector machine (SVM), which is used for the multi-category web pages classification. It uses a portion of the data to train the system and finds several support vectors that represent training data in web pages. SVM uses direct decision functions. The problem in support vector machine is that it considers classifying two values for given training data sets by hyper-plane of some feature space. Thus an extension to multiclass problem is not straightforward and the algorithm runtime for multi-category classification is overlong.

k Nearest Neighbor (kNN) method is often used in text document classification. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its *k* nearest neighbors. The accuracy of the *k*-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance. The *k*-nearest neighbor algorithm is sensitive to the local structure of the data. Web pages contain many irrelevant features and RF classifier can handle irrelevant feature without variable deletion and gives estimate of what variables are important in the classification.

In this paper, we described multi-category classification by using RF classifier. RF classifier can classify multi-category with many decision trees and provide reliable results on classification of web pages.

The rest of this paper is organized as follows. In section 2, related works are introduced. Section 3 describes about the Random Forest method. Proposed approach is explained at section 4 and experimental results are shown in section 5. Conclusion remarks are given in section 6.

## 2. Related Work

In this section, we give a brief survey of currently proposed approaches for web page classification. For Thai Academic Web Page

Classification , Verayuth Lertnattee and Thanaruk Theeramunkong [1] proposed inverse class frequency and web link information instead of inverse document frequency. There are two approaches that they represented for the problem of a large number of unique terms in a web pages (1) term weighting schemes and (2) schemes using Web link information. The simple term weighting of tf*idf is not good for classifying a small set of Web documents when they were categorized by the source of information by their preliminary experiments. Therefore inverse class frequency (icf) is used instead of inverse document frequency to alleviate the problem of classification of Thai academic web document. The icf was proved that it is useful in English collections which have a large number of terms [7].

For web information classification[2] , by applying binary decision tree into SVM method, an assorted DSVM method (Decision Support Vector Machine) was proposed [2]. SVM was developed from the optimum classification side under condition of linear separation. For text classification task, decision tree algorithm may select words that have information content based on information gain standard and can forecast text category which was affiliated with according to appearance condition of word combination in documents. SVMs classifying algorithm was only used in two values classification initially, lacks of the ability to deal with multi-value classification. In DSVM method, SVM classification algorithm was combined with basis of binary decision tree to form classifier of multi-category. DSVM has reduced the training sum of every SVM classifier greatly and has raised training efficiency.

Min-Yen Kan and Hoang Oanh Nguyen Thi analyzed [3] the usefulness of the uniform resource locator (URL) alone in performing web page classification. Their approach segmented the URL into meaning chunks and added component, sequential and orthographic features to model the salient pattern. In their proposed approach, they intended to concentrate on URL feature extraction and have added features to model URL component length, content, orthography, token sequence and precedence.

In the field of web page classification, Majid Javid Moayed et.al [4] investigated the usage of the swarm intelligence algorithm. Ant miner II was used for focusing on Persian web pages. They also proposed a simple text preprocessing technique to reduce the large number of attributes associated with web content mining. Their proposed preprocessing technique is efficient in the field of the web page classification. The Swarm Intelligence methods were identified with the feature of adaptability with the changes of environment. Also these methods have less computational costs than other methods like neural networks.

Oh-Woog Kwon and Jong-Hyeok Lee [5] proposed a Web page classifier based on an adaptation of k-Nearest Neighbor (KNN) approach. To improve the performance, they supplemented k-NN approach with a feature selection method and a term-weighting scheme using mark up tags and reform document-document similarity measure used in vector space model. And they also lacked to investigate the method that increases training samples using connectivity analysis in the WWW and then evaluate the proposed classifier using the extended training samples.

A web page classification based on support vector machine using a weighted vote schema for various features was described by Rung-Ching Chen and Chung-Hsun Hsieh [6]. This system used both Latent Semantic Analysis (LSA) and web page feature selection and recognition by the SVM model. Latent Semantic Analysis was used to find the semantic relations between keywords and between documents. The LSA classified semantically related web pages, offering users more complete information.

In this paper, we classify the web pages using RF classifier without using other feature selection methods. In our approach, web pages are classified for the user’s satisfaction and requirement efficiently and quickly.8

## 3. Random Forest Method

The Ensemble classification methods [9] train several classifiers and combine their results through a voting process. Random Forest method is the learning ensemble classification method consisting of a bagging of un-pruned decision tree learners with a randomized selection of features at each split. Random Forest method can also be used feature selection alone. In this paper, Random Forest method is investigated in multi-category classification of web pages.

The detailed steps of the random forest algorithm [10] are as follows:

Let *Ntrees* be the number of trees to build

for each of *Ntrees* iterations

**1**. Select a new bootstrap sample from training set

**2**. Grow an un-pruned tree on this bootstrap.

**3**. At each internal node, randomly select $m_{try}$ predictors and determine the best split using only these predictors.

**4**. Output overall prediction as the majority vote from all individually trained trees.

## 4. Proposed Approach

In this paper, we proposed Random Forest Classifier (RF) to classify web pages according to their corresponding categories. RF classifier is used when we have very large training data set and very large number of input variables. RF classifier based on the random forest method is not necessary to use other feature selection methods. Fig. 1 shows the general architecture of the proposed system.

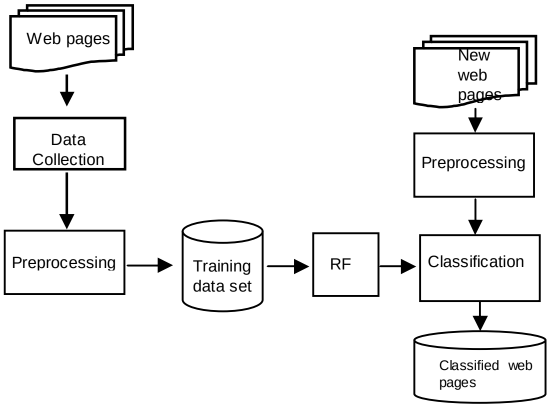


**Figure. 1 General architecture of the proposed approach**

## 4.1 Data Collection

We use the web pages from the Yahoo web site to collect training examples for our learning problem. Yahoo is best known for maintaining a web categorization directory. The web directory in Yahoo is a multilevel tree structure hierarchy. The top level of the tree which is the first level below the root of the tree. We use the top-level categories in Yahoo to label the web pages in our training data set. In our proposed system, we would like to classify the 5 categories of the web pages that categories are art & humanities, business & economy, computer & internet, health and sport.

We randomly selected over 3800 web pages from the Yahoo category. We downloaded 50 example web pages from each category. Table 1 shows the number of web pages that we used from the Yahoo for each category.

**Table 1. number of web pages from Yahoo web site**

| No | Category | Number of web pages |
|---|---|---|
| 1 | Entertainment & Art | 783 |
| 2 | Business & Finance | 997 |
| 3 | Sport & Recreation | 745 |
| 4 | Health & Wellness | 772 |
| 5 | Computer & Internet | 506 |
| | Total | 3803 |

## 4.2 Preprocessing

There are many kinds of information in the web pages such as text including noun and stop words such as articles, conjunctions etc. and images, link structure and layout. All of these are unnecessary to determine the class of the web page. Firstly, we deleted all of the tags such as <A></A>, <img></img>, <style></style>, <script></script> and advertisement banner, navigation bar and copy right notice from each web pages. And then removed all of the stop words according to the existing standard stop word list. Some technical term consisted in WWW's field are insensitive and assumed that are worthless item. Thus, we put such words to a stop list.

After removing all of the unnecessary tags and words from the web pages, we analyzed through the web pages to get the important attributes for classifying the corresponding web pages. In the analysis of the study, we found that some of the terms are always exist in the web pages for the same category. We used these terms as attributes for classification of web pages. We derived the training data set from these terms.

## 4.3 Random Forest Classifier (RFC)

RFC is a collection of individual decision tree classifiers, therefore it can improve classification accuracy. In our system, each tree is constructed using a different bootstrap sample of the training dataset. For each observation, each individual tree votes for one class and the RFC classifies the class that has the plurality of votes. Fig.2 shows the general architecture of RFC.

Firstly, RFC bootstrapping the training data set $D_x$. Assume that there are N training cases, $D_x = \{t_1, t_2, ..., t_N\}$. After getting the bootstrap, it became the new training set $E_x$.

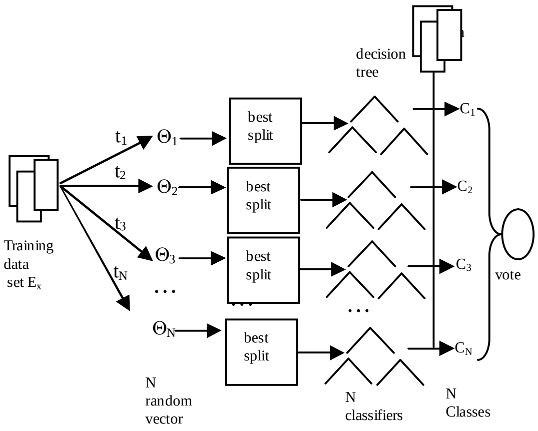


**Figure 2. General architecture of RF classifier.**

$E_x = \{t_1, t_1, t_3, ..., t_N\}$. In each training case, there consists of M attributes, $t_i = \{attr_1, attr_2, ..., attr_M\}$. Construct the decision trees for each training case in $E_x$. To construct the decision trees, RFC randomly selected the attributes $m_{try}$ from the training case. The number of attributes ($m_{try}$) is always less than the number of attributes M in the training case. According to the basic concept of random forest $m_{try} = \sqrt{M}$.

After getting $m_{try}$ attributes, choose the most important attributes among them to construct the decision trees. To choose the important attributes for best split of the tree used gini index.

$$gini(attr) = 1 - \sum [Pj]^2$$

$$gini_{split} = \sum_{attr=1}^{m_{try}} \frac{n_{attr}}{n} gini(attr)$$

where $P_j$ is the relative frequency of the attribute (attr) at class j , $m_{try}$ is number of attributes, $n_{attr}$ is the number of randomly selected training records, n is the total number of training records.

The root node of the tree is the attribute with lowest $gini_{split}$. Decision trees are constructed according to the $gini_{split}$ of each attribute. The most important attribute is based on the minimum value of $gini_{split}$. Randomly constructed trees are fully grown and not pruned. RFC uses the voting process to choose the most popular class.

In this paper, number of decision tree is limited to the number of training case. Each tree produces one class as output and RFC consists of ensemble of tree. RFC votes the most popular class as classified web pages.

## 5. Experiments

The data used in the experiment were collected from Yahoo web site. To test the proposed system, we collected web pages on 5 domains of 14 categories in Yahoo web site, “Art & Humanities”, “Business & Economy”, “Sport”, “Computer & Internet”, and “Health”. We randomly downloaded 50 web pages per category and the total number of web pages for experiment is 250.

We analyzed all of the downloaded web pages. After analyzing the web pages, we found that some of the terms are frequently contained in the same web pages. We may decide the class of the web pages according to these terms. Table (2) shows the number of attributes that we got from the analysis study of the web pages.

**Table 2. number of attributes**

| no | Class | no of attributes |
|---|---|---|
| 1 | Entertainment & Art | 97 |
| 2 | Business & Finance | 68 |
| 3 | Sport & Recreation | 143 |
| 4 | Health & Wellness | 160 |
| 5 | Computer & Internet | 78 |
| 6 | Total | 546 |

From these attributes, we got the training data set for classification of web pages. And then we constructed the RFC for each class using the training data set. In this paper, the classification accuracy of RFC is evaluated by error rate to compare the accuracy of the decision tree.

### 5.1 Performance Measure

The performance of the classifier in this paper is evaluated in terms of error rate to get accuracy. We defined error rate:

$$Errorrrate = \frac{noof\ all\ testing\ examples\ classified\ erroneously}{noof\ all\ testing\ examples}$$

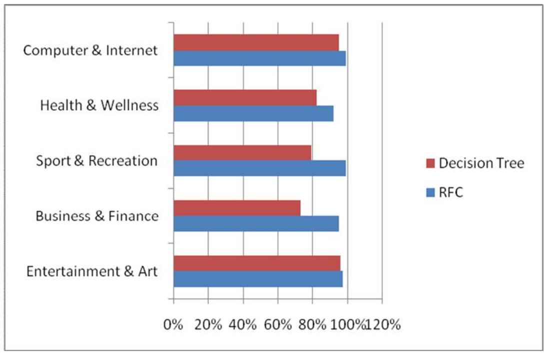
We compared the accuracy of our system with the accuracy of decision tree classifier for each category. In the next section, we will be shown experimental results compared with decision tree classifier. The results show that our system can classify the web pages more accurately than the decision tree classifier.

## 6. Experimental Results

From the testing of our system, we experimented that the accuracy of RFC in Entertainment & Art is 97%, Business & Finance is 95%, Sport & Recreation is 99%, Health & Wellness is 92% and Computer & Internet is 95 % and that of the decision tree 96%, 73%, 79%, 82% and 95% respectively. According to these results, we found that our system can classify more accurately than the decision tree classifier. Table (3) shows the comparison result of two classifiers.

| | Entertainment & Art | Business &Finance | Sport &Recreation | Health &Wellness | Computer&Internet |
|---|---|---|---|---|---|
| RFC | 97% | 95% | 99% | 92% | 99% |
| Decision Tree | 96% | 73% | 79% | 82% | 95% |

The following figure, Fig. 3 is the bar chart of the comparison results.



## 7. Conclusion

In this paper, we represented the effect of RFC approach for web page classification. The characteristics of this approach is firstly applied morphological analysis to web pages and extract noun then using basket analysis generated attributes. And then using RFC to classify web pages. Compared with decision tree classifier, our experiments have shown that RFC based classification method performed better in error rate so classification accuracy is higher than the decision tree classifier but there is no significance difference in precision and recall. Our approach performed in better accuracy than decision tree classifier on the same data set.

## References

[1] V. Lertnattee and T. Theeramunkong, "Improving Thai Academic Web Classification Using Inverse Class Frequency and Web Link Information", 22nd International Conference on Advanced Information Networking and Application-Workshops, 2008.

[2] L.Liu et.al, "The Research of Decision Support Vector Machine in Web Information Classification", 12th International Conference on Computer Supported Cooperative Work in Design, pp. 196-200, 16-18 April, 2008.

[3] M. Yen Kan and H. Oanh Nguyen Thi, "Fast Webpage Classification Using URL Features", *ACM*, Berman, Germany, November 2005.

[4] M. Javid Moayed et.al, "Ant Colony Algorithm for Web Page Classification", International Symposium on Information Technology, 2008.

[5] O. Woog Kwon and J. Hyeok Lee, "Web Page Classification Based on k-Nearest Neighbor Approach", in the proceedings of fifth International workshop on Information retrieval with Asian Languages, Hong Kong, China, pp. 9-15, 2000.

[6] R. Ching Chen and C. Hsun Hsieh, "Web Page Classification based on a support vector machine using a weighted vote schema", Expert Systems with Application, Volume 31, Issue 2, pp. 427-433, August 2005.

[7] L. Liu, H. Song, "Combining Fuzzy Clustering with Naïve Bayes Augmented Learning in Text Classification", The First International Symposium on Pervasive Computing and Applications, pp. 168-171, 2006.

[8] http://www.stat.berkeley.edu/~breiman/Random Forests/cc_home.htm.
K.Remlinger, "Random Forest as a variable selection Toll for Biomarker Data". ICSA Applied Statistics Symposium, June , 2007.

[9] S.R.Joelsson, J. Atli Benediktsson and J.R.Sveinsson, "Random Forest Classifier for Hyperspectral Data".

[10] A.A.Montillo, "Random Forest". Guest lecture: Statistical Foundation of Data Analysis, February, 2009.