Approach for Dimensionality Reduction in Web Page Classification

Shraddha Sarode
Computer Engineering (M.E),
Thadomal Shahani Engg. College,
Mumbai, India

Jayant Gadge
Computer Engineering (M.E),
Thadomal Shahani Engg. College,
Mumbai, India

ABSTRACT
Dimensionality refers to number of terms in a web page. While classifying web pages high dimensionality of web pages causes problem. The main objective of reducing dimensionality of web pages is improving the performance of classifier. Processing time and accuracy are two parameters which influence the performance of a classifier. To reduce the processing time, less informative and redundant terms have to be removed from web pages.

This research describes hybrid approach for dimensionality reduction in web page classification using a rough set and naïve Bayesian method. Feature selection and dimensionality reduction methods are used for reducing the dimensionality. Information gain method is used as feature selection method. Rough set based Quick Reduct algorithm is used for dimensionality reduction. Naïve Bayesian method is used for classifying web pages to optimal predefined categories. Assignment of web pages to category is based on maximum posterior probability. Words remaining after the process of feature selection and dimensionality reduction will be given to the classifier. Finally the classifier will assign most optimal predefined category to web pages.

Keywords
Dimensionality Reduction, Feature Selection, Information gain, Naïve Bayes, Rough Set, Web Page Classification.

1. INTRODUCTION
Today there is huge amount of data available on World Wide Web. While searching for any document on the internet, many irrelevant results are found. It is time consuming and frustrating to keep on searching until desired results are obtained. Therefore, web page classification can be used to provide fast and efficient results to users. Classifying such a huge information manually is time consuming and costly. Therefore, Web page classification task automation of is very helpful for improving the results of many applications for example information retrieval.

Classification is a supervised learning method. Web page classification can be defined as a process of assigning web pages to the most optimal categories [1]. Web page classification can be used in various applications ranging from email monitoring [2] to medical diagnosis. There are many applications of web page classification and some are web content filtering, ontology annotation [2], assisted web browsing contextual advertising and knowledge base construction, constructing, maintaining or expanding web directories (web hierarchies), helping question answering systems, building efficient focused crawlers or vertical (domain-specific) search engines, improving quality of search results [3]. Web page classification also plays main role in information retrieval.

High dimensionality is the major problem in web page classification because amount of data is increasing rapidly on www. For example if a training dataset contains 500 web pages in each predefined category for those web pages. Then for training classifier, selection of relevant terms becomes an issue. Therefore, feature selection and dimensionality reduction are used to remove terms that are less informative, redundant and irrelevant. Feature selection technique solves the problem of selecting the input features that are most predictive for given predefined categories. These two methods will be useful to improve the classification accuracy and efficiency.

Information gain method and rough set method is used for feature selection and dimensionality reduction respectively. Information gain gives terms that are the most informative in assigning web pages to categories. Rough set based supervised quick reduct algorithm uses dependency measure for dimensionality reduction. Dependence measure is also known as correlation measure. It quantifies the ability of feature to predict the value of decision attribute from the value of conditional attributes. In this paper, conditional attributes are words from the web pages and decision attribute is predefined category of web page. Reduced terms from quick reduct algorithm will be the input for naïve Bayesian classifier.

For web page classification, various machine learning algorithms have been utilized. In proposed system, Naïve Bayesian method is used for classifying web pages. Naïve Bayes (NB) is one of the popular machine learning algorithms because of its simplicity and fast learning in case of large datasets. Simplicity lies in classification based on calculation of probability by the naïve Bayes independence assumption [4].

The paper is organized as follows. Section 2 presents related work. The details of methods used in the proposed approach are described in section 3. In the section 4 steps followed in proposed system are mentioned. Results for feature selection and dimensionality reduction are combined and shown in section 5. Section 6 explains different measures used to evaluate the performance of a classifier. In the end, section 7 presents conclusion.

2. RELATED WORK
Juan Zhang, Yi Niu, Huabei Nie (2009) combined fuzzy and k-nearest neighbor algorithm. The k- nearest neighbor is a simple classification algorithm that is used to assign patterns of unknown classification to the class of the majority of its k nearest neighbors of known classification based on the distance measure, and drawback of the method is that each of the patterns of known classification is considered equally important in the assignment of the pattern to be classified. In fuzzy K-NN method instead of giving equal importance to each pattern, assign class membership as a function of the

pattern distance from its k-nearest neighbors and those neighbors' memberships in the possible classes. For feature selection term frequency inverse document frequency (tfidf) method is adopted. Web document preprocessing has steps as reducing morphological variants of words to a root form, pruning of infrequent words, Pruning of high frequent words such as 'the', 'a', 'an' etc. [5]

Rung-Ching Chen, Chung-Hsun Hsieh (2006) proposed method for web page classification based on a support vector machine using a weighted vote schema. [6] It is a method for both linear and nonlinear data. It uses a nonlinear mapping to transform the original training data into a higher dimension. With the new dimension, it searches for the linear optimal separating hyper plane (i.e., "decision boundary"). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyper plane. Preprocessing includes removal of html tags and Chinese word segmentation. In Latent semantic analysis, svd is applied to decompose the matrix and the original data vectors are reduced to a small number of features. Web page feature selected based on four measures that are the number of keywords from term database, the number of words in a document, the ratio between the number of keywords and the number of words, and the average interval between each term. In classification semantic features and text features to train the SVM. The two SVM category models are used to predict the category of the web pages. After the two SVM models classify the web pages, voting policy used to vote to which category the web page should be assigned. [6]

Hybrid approach of Rough set and Genetic Algorithm for web page classification is proposed by Xiaoyue Wang, Zhen Hua and Rujiang Bai (2012). Rough set is used for dimensionality reduction. GA is based on an analogy to biological evolution. An initial population is created consisting of randomly generated rules. Based on the notion of survival of the fittest, a new population is formed to consist of the fittest rules and their offspring. The fitness of a rule is represented by its classification accuracy on a set of training examples. Offspring are generated by crossover and mutation. The process continues until a population P evolves when each rule in P satisfies a prespecified threshold. It is adaptive, robust, efficient, and global search method, suitable in situations where the search space is large. Optimization of a fitness function is according to the preference criterion. Before fitness evaluation authors have used support vector machine classifier for training and testing document features that is for dimensionality reduction which provides three mentioned advantages, to reduce the high dimension of feature vectors, to set the best kernel parameters and to choose the optimal input feature subset for SVM. [7]

One of the variant of Naïve Bayesian technique is introduced by G.S. Tomar, Shekhar Verma, and Ashish Jha (2006). In this paper, preprocessing steps involved are Hypertext filtering, tokenizing and Case conversion, Stemming. The proposed approach for classification of web text is based on the concept of modified naïve Bayesian method, instead of taking each word for classification; only the relevant words are taken into account. The relevancy is calculated using the word weighting scheme which is based on term frequency inverse document frequency (tfidf). Advantage of this modified Naïve Bayesian algorithm is based on relevant words consideration increases accuracy of classification and reduces computational time of calculating word probabilities [8].

As each term in HTML tag for each web page can be taken as a feature. It causes the problem of high dimensionality. To reduce dimensionality problem, Selma Ayse Özel (2009) proposed optimal feature selection technique based on Genetic algorithm. The performance of this method is compared with J48 (decision tree), the Naïve Bayes Multinominal (Bayes), and the IBk (kNN) classifiers. It gives 96% accuracy using GA as feature selector. In this method, the numbers of features considered are large i.e. up to 50000 features, system takes both terms and HTML tags together on a Web page as features, assign different weights to each feature and the weights are determined by the GA. After extracting features, document vectors for the Web pages are created by counting the occurrences of each feature in the associated HTML tag of each Web page. The GA feature selector consists of coding, generation of initial population, evaluation of a population, reproduction, crossover, mutation, and determination of the new generation steps and reproduction, crossover, mutation steps are repeated, number of generations times until optimal feature vector found [9].

3. THEORETICAL BACKGROUND

In this section, the basic concepts behind the techniques used in the proposed system are discussed.

3.1 Web Page Classification

Web page classification is a process of assigning web pages to optimal categories. Contents of web pages are used to predict the category of web pages. One major problem in using contents for web page classification is high dimensionality. To address this problem, only relevant features and important features or terms should be taken into account rather than considering all features. Feature is relevant if the feature is capable of prediction of the decision feature. For addressing the problem of high dimensionality, feature selection and dimensionality reduction methods are adopted.

3.2 Information Gain

Feature selection has been a field of research and development since 1970s in machine learning [10]. This method is adopted for feature selection to remove less informative words.

Information gain is calculated for each term. Threshold is set for removing less informative words. Terms with values less than predefined threshold are removed. This will help in getting more informative words rather than considering all terms [11].

The information gain computation involves the calculation of entropy and conditional probabilities of category given term. Formula for information gain is given as follows:

$IG(t) = -M_1+M_2+M_3$ … (1)

Where,

$$M_1 = \sum_{i=1}^{n} P(C_i) log\, P(C_i) \quad \dots (2)$$

$$M_2 = P(t) \sum_{i=1}^{n} P(C_i|t) log\, P(C_i|t) \quad \dots (3)$$

$$M_3 = P(\bar{t}) \sum_{i=1}^{n} P(C_i|\bar{t}) log\, P(C_i|\bar{t}) \quad \dots (4)$$

Where t denotes term, C_i denotes the i^{th} category, m denotes the number of categories, $P(Ci|\bar{t})$ and $P(Ci|t)$ denotes conditional probabilities of category given that term appears in document and given that term does not appear in the document respectively.

3.3 Rough Set Based Quick Reduct Algorithm

Rough set theory was proposed by the author Zdzisław Pawlak in 1982. Rough Set method is used for dimensionality reduction. Information system is the main concept in rough set. Information system is a table that contains conditional attributes and decision attribute. Rough set can be defined by two operations known as approximations [12].

Suppose there are objects in U. It has a subset X. R ⊆ U × U is an equivalence relation. But there is lack of knowledge about elements of U. For classification of set X with respect R approximation will be used. Lower approximation is a set of objects which are to be surely classified as elements of set X. Upper approximation is set of objects which may possibly classify as elements of set X. Lower approximation and upper approximation are also known as positive region and negative region respectively. Boundary region of X is the difference between upper approximation and lower approximation. For a set to be rough boundary region it should not be empty. Rough set is a method to discover dependency between features and reducing those features. Reducing features is the process of finding a set of attributes which is called reduct. Reduct is the ratio of cardinality of positive region of set to the cardinality of universal set.

Only features which are the most informative are kept. For dimensionality reduction Zdzisław Pawlak proposed QuickReduct Algorithm. Supervised QuickReduct Algorithm is used in this research.

Dependency measure in quick reduct algorithm is calculated using following formula:

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|U|} \quad \ldots.(5)$$

Where P is a set of condition attributes and Q is the decision attribute, $\gamma_p(Q)$ is the dependency between condition attributes and decision attribute. If k=1, Q depends totally on P; if 0<k<1, Q depends partially on P, and if k=0 then Q does not depend on P. The goal of attribute reduction is to remove redundant attributes. Reduced set of condition attributes provides the same quality of prediction as the original attributes.

QuickReduct Algorithm is shown below:
C: the set of all conditional attributes,
D: the set of decision attributes

QuickReduct (C, D)

- Algorithm starts with empty set R – {}
- It adds one attribute at a time T ←R
- For each attribute dependency is calculated with respect to decision attribute and compared. $\gamma_{RU\{x\}}(D) > \gamma_T(D)$
- If it provides the greatest increase in dependency metric then it gets add in the reduced set T ←R U{x}
- This process continues until algorithm produces its maximum possible value for attributes of the dataset. $\gamma_R(D) == \gamma_C(D)$

Algorithm produces an output which is a reduced set of conditional attributes. The main advantage of using rough set theory based supervised QuickReduct algorithm is that it does not need any additional information about data for finding minimal set of attributes. [13]

3.4 Naïve Bayesian Classifier

Naïve Bayesian classifier is a statistical classifier. It is based on Bayes' theorem. [14] It calculates posterior probability that particular hypothesis holds for given data. Posterior probability is calculated using prior probability of hypothesis, likelihood of data and observation of data or evidence. Naïve Bayesian is most commonly used technique for classification because of its simplicity. This algorithm makes assumption of independence between each pair of words.

Therefore, this works well in domains with many equally important features. Although it makes an unrealistic assumption of independence, this method is exceptionally successful with large datasets. [15] In web page classification, the number of features or terms can easily rise to hundreds and thousands and that is also known as problem of high dimensionality. The big hurdle in applying many sophisticated learning algorithms to web page classification is to decide how to represent the arbitrary text document in terms of attribute values.

In the context of web page classification naïve Bayesian algorithm assigns web pages to the most optimal predefined category. In this research, naïve Bayesian classifier calculates conditional probabilities for each term against predefined category. Optimal category is assigned to web pages based on maximum posteriori probability considering the term frequencies. Therefore, this method is known as probabilistic method.

This method is incremental in nature because each training example can incrementally increase or decrease the probability of a hypothesis being correct. In web page classification, hypothesis means whether web page belongs to a category or not [16].

For assigning web page to category following formula is calculated for each document for every category [17]:

$$P(c|document) = P(c) * P(word_1|c) * P(word_2|c) * \ldots * P(word_n|c) \quad \ldots.(6)$$

Where, c denotes category, and document contains n words.

For calculating conditional probabilities P(word|c) use of laplacian correction is required. The reason for using laplacian correction is to prevent zero probabilities for terms which are not present in training examples. Formula for P(word|c) using laplacian correction is given as follows:

$$P(word|c) = \frac{1 + count\ (w, c)}{|V| + count\ (c)} \quad \ldots.(7)$$

Where, |V| denotes number of distinct words in all training examples, count(c) denotes total number of words in category c, count (w, c) contains occurrence of w in c.

And finally for each document, $P(c_1|document)$, P(c2|document),…., $P(c_m|document)$ is compared. Document is assigned to category based on maximum value of P(c|document) calculated above which nothing but posterior probability.

$$ArgMax_{cj} \in c P(c_j | d_i) \approx ArgMax_{cj} \in c \prod_{k=1}^{n} P(t_{ik} | c_j) P(c_j) \ldots..(8)$$

4. PROPOSED SYSTEM

The proposed architecture includes three parts. Document preprocessing which is the first process for web page classification. Information Gain method is used for feature selection. Rough set based supervised quick reduct algorithm is used for dimensionality reduction method. Naïve Bayesian classifier is applied after dimensionality reduction to documents for getting classified documents.

4.1 Preprocessing step includes:

- Tokenizing, tokenize the file into individual tokens using space as the delimiter.
- Stemming, the words are converted to their root form. This is achieved using Porter's stemming algorithm.
- Stop words removal, removing all words which do not convey any meaning.

4.2 Feature Selection and Dimensionality reduction step includes:

- Calculation of information gain for each term [8].
- Features are selected based on high information gain values.
- Construction of information system with conditional and decision attributes.
- Supervised quick reduct Algorithm is used for dimensionality reduction which will give reduced features.
- Calculation of output of quick reduct algorithm using dependency measure.

4.3 Naïve Bayesian classifier step includes:

- Calculation of word probabilities against predefined categories considering term frequency in documents and using laplacian correction.
- All the word probabilities for each category are summed up.
- Calculation of maximum posterior probability.
- Assignment of web pages to most optimal predefined category based on maximum posterior probability.

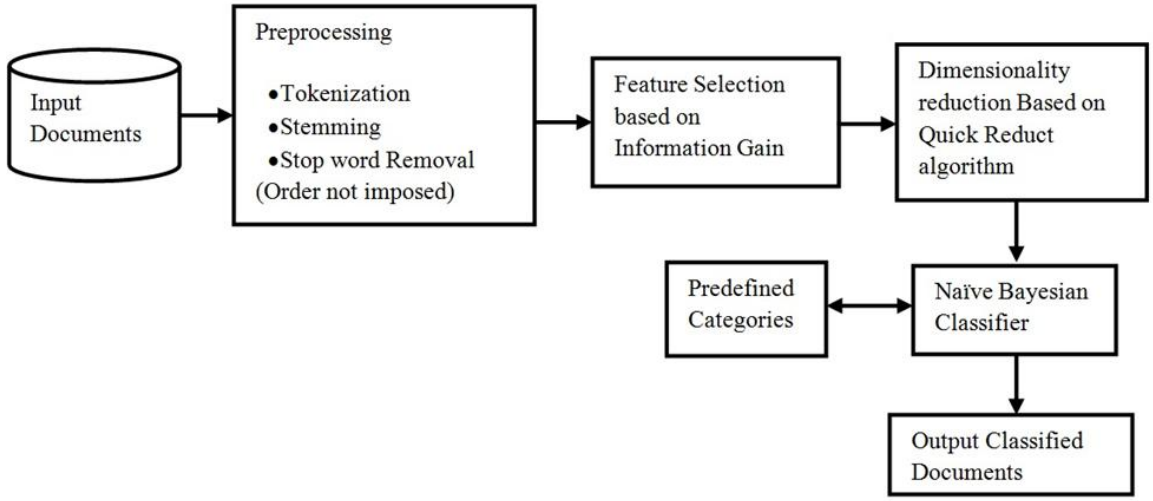

Fig 1: Proposed System for Web Page Classification

5. RESULTS

Results after applying information gain as feature selection method and rough set algorithm as dimensionality reduction method are shown in Table 1. Reduced number of words will be given as input to Naïve Bayesian classifier. From observation of following table one can conclude that definitely the processing time will be less for reduced number of words as compared to total number of words.

Table 1: Result of Dimensionality Reduction

Number of categories	Total number of documents	Total number of words	Reduced number of words
2	**678**	**12651**	**1329**
4	**1396**	**32980**	**5971**
8	**5485**	**71671**	**9996**
20	**7528**	**85116**	**34305**

6. EVALUATION

Processing time and accuracy are two parameters which influence the performance of classifier. Processing time refers to the time duration required to complete the task of classification after dimensionality reduction process. Processing time of the classifier can be reduced using dimensionality reduction method. Because of the number of terms processed by a classifier is reduced. A confusion matrix contains information about actual and predicted classifications. Classifier accuracy is evaluated using the data in the confusion matrix. This is illustrated by the Table 1 shown below.

- True Positive refers to the number of documents correctly classified to that category.
- True Negative refers to the number of documents correctly rejected from that category.
- False Positive refers to the number of documents incorrectly rejected from that category.
- False Negative refers to the number of documents incorrectly classified to that category.

The precision, recall and F1 measure are used as evaluation measures and are calculated using the following formulae:

Precision is the probability of a correctly rejecting document from that category; it is also known as true negative rate or specificity.

$$precision = \frac{TP}{TN + FP} \quad \ldots.(9)$$

Recall is the probability of a correctly classifying documents and assigning to optimal category, it is also known as sensitivity or positive predictive value.

$$recall = \frac{TP}{TN + FN} \quad \ldots.(10)$$

The F_1 score interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst score at 0.

$$F1\ measure = 2\left(\frac{P * R}{P + R}\right) \quad \ldots.(11)$$

Table 2: **Confusion Matrix**

	Predicted class	
Actual class	TP (true positive)	FP (false positive)
	FN (false negative)	TN (true negative)

As dataset contains are more than two categories, Micro averaging Precision, Micro averaging Recall and Micro averaging F1 measures are used for evaluation [18].

$$\text{Microaveraging precision } (\pi^{\mu}) = \frac{\sum_{i=0}^{|C|} TP_i}{\sum_{i=0}^{|C|}(TP_i + FP_i)} \ldots.(13)$$

$$\text{Microaveraging recall } (\rho^{\mu}) = \frac{\sum_{i=0}^{|C|} TP_i}{\sum_{i=0}^{|C|}(TP_i + FN_i)} \ldots(14)$$

$$\text{Microaveraging F1 } (F1^{\mu}) = 2 * (\frac{P*R}{P+R}) \ldots.(15)$$

7. CONCLUSION

As there are billions of web pages on internet, management of these web pages is essential. Web page classification is difficult task because of different size and formats of web pages. This paper proposes hybrid approach for dimensionality reduction in web page classification. Less informative and redundant terms are removed using feature selection and dimensionality reduction methods.

Feature selection and dimensionality reduction methods overcome the problem of high dimensionality. The proposed approach would improve the accuracy and efficiency of classifier. Because of dimensionality reduction, this approach would save processing time. As compared to the traditional approach, this method requires less processing time because it uses dimensionality reduction technique.

8. REFERENCES

[1] Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining: Concepts and Techniques, 3rd Ed., Han, Kamber & Pei, University of Illinois at Urbana-Champaign &Simon Fraser University, 2011

[2] Ming Mao, Yefei Peng, Michael Spring, "Ontology Mapping: As a Binary Classification Problem", IEEE Fourth international conference on Semantics, Knowledge and grid, 2008

[3] Xiaoguang Qi and Brian D. Davison, "Web Page Classification: Features and Algorithms", *ACM Computing Surveys*, Vol. 41, No. 2, Article 12, Publication date: February 2009.

[4] Tom M. Mitchell, "Machine Learning," Carnegie Mellon University, McGraw-Hill Book Co, 1997.

[5] Juan Zhang, Yi Niu, Huabei Nie, "Web Document Classification Based on Fuzzy k-NN Algorithm", *International Conference on Computational Intelligence and Security*, IEEE, 2009.

[6] Rung-Ching Chen *, Chung-Hsun Hsieh, "Web page classification based on a support vector machine using a weighted vote schema", *Expert Systems with Applications* 31, Elsevier, 2006.

[7] Xiaoyue Wang, Zhen Hua, Rujiang Bai. "A Hybrid Text Classification model based on Rough Sets and Genetic Algorithms" *Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, IEEE, 2012.

[8] G.S. Tomar, Shekhar Verma, Ashish Jha, "Web Page Classification using Modified Naïve Bayesian Approach", *IEEE*, 2006.

[9] Selma Ayse Özel, "A Genetic Algorithm Based Optimal Feature Selection for Web Page Classification", *IEEE*, 2011.

[10] Tseng, V.S.; Ja-Hwung Su; Hao-Hua Ku; Bo-Wen Wang;" Intelligent Concept-Oriented and Content-Based Image Retrieval by using data mining and query decomposition techniques" IEEE International Conference on Multimedia and Expo. June 23 2008-April 26 2008 Page(s):1273 – 1276

[11] Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." In *ICML*, vol. 97, pp. 412-420. 1997.

[12] Pawlak, Zdzisław. "Rough sets." International Journal of Computer & Information Sciences 11.5 (1982): 341-356.

[13] C. Velayutham and K. Thangavel, "Improved Rough Set Algorithms for Optimal Attribute Reduct", *Journal of Electronic Science and Technology*, VOL. 9, NO. 2, June 2011

[14] Ramez Elmasri, Shamkant B. Navathe, "Fundamentals of Database Systems,"Addison Wesley Longman Publishing Co., Fifth Edition, 2007.

[15] Sang-Bum Kim, Kyong-soo Han, Hae-Chang Rim, Sung Hyon Myaeng "Some Effective techniques for Naïve Bayes Text Classification" *IEEE Transactions on Knowledge and Data Engineering* -2006

[16] Vidhya.K.A, and G.Aghila, "Hybrid Text Mining Model for Document Classification", *The 2nd International Conference on Computer and Automation Engineering (ICCAE)*, 2010.

[17] Dino Isa, Lam Hong Lee, V.P. Kallimani, and R. RajKumar, " Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine", *IEEE transactions on knowledge and data engineering*, vol. 20, no. 9, September 2008

[18] Franca Debole & Fabrizio Sebastiani, "An Analysis of the Relative Hardness of Reuters-21578 Subsets", Technical report, *Institute of Science and Technologies of the National Research Council Via Giuseppe Moruzzi*, Pisa, Italy, 2003