ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

Classification of Web Pages Using Naive Bayesian Approach

Bhagyesh P. Asatkar[1], Prof. K. P. Wagh[2], Dr. P N. Chatur[3]

PG Scholar, Department of Computer Science and Engineering, Government College of Engineering, Amravati, India[1]

Assistant Professor, Department of Information Technology, Government College of Engineering, Amravati, India[2]

Associate Professor, Department of Computer Science and Engineering, Government College of Engineering, Amravati, India[3]

ABSTRACT:In World Wide Web there is huge amount of data available and the number is increasing. Every day millions of web pages being created and uploaded in the electronic form. So, one way to manage this large amount of data is to classify this data. Classification is the process that divides data into groups, which are either dependent or independent of each other and each class is acts as a class. High dimensionality is one of the issues in the web page classification. Dimensionality is nothing but the number of terms in a web page. High dimensionality of web pages causes problem while classifying them. The main objective of reducing dimensionality of web pages is to improve the performance of the classifier. In this survey paper, we describes a feature selection and dimensionality reduction to reduce the dimensionality with the help of information gain and rough set based Quick Reduct algorithm respectively and then with the help of Naive Bayesian classifier web pages are classified.

KEYWORDS: Dimensionality Reduction, Information Gain, Feature Selection, Naive Bayesian Classifier.

I. INTRODUCTION

Text Categorization is an automatic technique use to label the documents with a predefined set of topics over the wide availability of web documents in electronic forms. Learning in a very high dimensional data space is the key challenge in text categorization. For a given data set, a collection of all words or phrases forms a “dictionary” with hundreds of thousands features, each word or phrase occurs in documents once or more times is consider as a feature.High dimensionality of features leads to the high computational burden and may affect the performance of classifier due to redundant and irrelevant features. To overcome the curse of dimensionality and to fasten the learning process of classifier, there is a need of feature reduction to reduce the size of feature. In feature selection approach,only feature subset is use and the rest of them get discarded. In the process of web page classification, there are issues such as high dimensionality, mining a knowledge from diverse data, incorporation of background knowledge, handling noise and incomplete data. Therefore, for removing the terms that are redundant, irrelevant and less informative, use feature selection and dimensionality reduction technique. These techniques solve the problem of selecting the input feature that is most informative and relevant to improve the classification accuracy and efficiency.For feature selection and dimensionality reduction, Information Gain and rough set based Quick Reduct algorithm used.For assigning web pages to predefined categories, information gain gives terms that are the most informative and rough set based supervised Quick Reduct algorithm uses dependency measure to reduce dimensionality. Dependence measure is also known as correlation measure. Itquantifies the ability of a feature to predict the value of decisionattribute using values of conditional attributes. Conditional attributes are words from the web pages and decisionattribute is predefined category of web page. Reduced terms fromquick reduct algorithm will be given as a input for Naïve Bayesian classifier.Naïve Bayesian method isused for classifying web pages. Naïve Bayes (NB) is one of thepopular machine learning algorithms. It is applicable in case of largedatasets because of its simplicity and fast learning. Its simplicity liesin Bayes assumption of independence between terms.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

The paper is organized as follows. Section II presents related work. The details of methods used for classification described in section III. The section IV presents system architecture and conclusion in sectionV.

II. RELATED WORK

For classification Juan Zhang, Yi niu, HuabeiNie[3] combines fuzzy and k-nearest neighbour algorithm. The k-nearest neighbour is a simple classification algorithm used for assigning patterns of unknown classification to the class of majority of its k-nearest neighbour of known classification based on the distance measure but the single drawback is that method gives equal importance to the patterns of known classification and the patterns to be classify.In fuzzy k-nearest neighbour method instead of giving equal importance to each pattern, assign class membership as a function of the pattern distance from its k-nearest neighbours and those neighbours' memberships in the possible classes. For feature selection term frequency inverse document frequency (tf-idf), method is used. Web document pre-processing 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.

Rung-Ching Chen, Chung-Hsun Hsieh[4] proposed method for web page classification based on a support vector machine(SVM) using a weighted vote schema. This method is for both linear and nonlinear data. It uses a non-linear mapping to transform the original training data into a higher dimension.The feature vectors are extracting from both the Latent Semantic Analysis(LSA) and Web Page Features Selection(WPFS) methods. The LSA can extract common semantic relations between terms and documents then it classifies semantically related web pages, offering users more complete information. The WPFS extracts four text features from the web page content. The web page category can correctly determine by the WPFS. They also compared the SVM performance using four different kernel functions performance. The experimental results show that thekernel function yields the best result of these four kernel functions. The LSA-SVM, BPN (Back Propagation Network)) and WVSVM(Weighted Voting Support Vector Machine) were then compared. The experiment demonstrated that the WVSVM gives better accuracy even with a small dataset. When the smaller category has less training data, the WVSVM is still able to categorize web pages with acceptable accuracy.

Sneha V. Dehankar, K.P.Wagh[9] proposed a system, on the basis of highest probability of word of each document various links are classified into predefined classes. Due to the apriori algorithm from all the links only several and relevant links are being classified. These classified reduced links helps to reduce the complexity and time to scan all the links.The web page classification system gives the F-Measure value of 75.91% and gives the efficiency and accuracy of classifier. The systems performance increases whenever keyword submitted to the systembecause the value of true positive parameter foreach keyword submitted is better while the value of false positive isless. The number of predefined class label and number of words that are predefine for each class can be increasefor thebetter accuracy in future.

ShraddhaSarode and JayantGadge[2]proposed a hybrid approach of dimensionality reduction for web page classificationusing a rough set andinformation gain method rather than giving all words to classifier, only informative and relevant words are given to the classifier. Less informative and redundant terms removed using feature selection and dimensionality reduction methods. From experimental results, we find that more than 60% less informative words are removed. To reduce the dimensionality of webpagesFeature selection and dimensionality reduction methods are used. Information gain method and Rough set based Quick Reduct algorithm used for feature selectionand for dimensionality reduction respectively. Web pages are classified using NaïveBayesian method.

Bo Tang, HaiboHe,Paul M. Baggenstoss and Steven Kay[1]presented a Bayesian classification approach for automatic text categorization using class-specific features. In contrast to the conventional feature selection methods, it allows to choose the most important features for each class. To apply the class specific features for classification, they derived a new naive Bayes rule following Baggenstoss's PDF Projection Theorem. One important advantage of this method is that many existing feature selection criteria can easily incorporated.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

III. BACKGROUND

In this section, we will discuss the basic concepts used for classification. Web page classification is a process of assigning web pages to optimal categories. To predict the category of web page, contents of web page are used. One major problem in using contents for web page classification is high dimensionality. To overcome this problem, only relevant features and important features or terms should take into account rather than considering all features. Feature is relevant only if it is capable of prediction of the decision attribute. For addressing the problem of high dimensionality, feature selection and dimensionality reduction methods are used. Information Gain and Rough Set based Quick Reduct methods are used as dimensionality reduction methods and Naïve Bayesian classifier is used for classification

.

Information Gain

For removing less informative words information gain method is used. Information gain is calculated for each term and threshold is set for removing less informative words. For getting, more informative wordsterms with values less than predefined threshold are removed. For computation of information gainequation1 involves the calculation of entropy equation 2 and conditional probabilities of category against term equation 3 and equation 4. Formula for information gain is given as follows:

$$IG(t) = -X_1 + X_2 + X_3 \quad (1)$$

Where,

$$X_1 = \sum_{i=1}^{n} P(C_i) \log P(C_i) \quad (2)$$

$$X_2 = P(t) \sum_{i=1}^{n} P(C_i|t) \log P(C_i|t) \quad (3)$$

$$X_3 = P(\bar{t}) \sum_{i=1}^{n} P(C_i|\bar{t}) \log P(C_i|\bar{t}) \quad (4)$$

Where C_i denotes the i^{th} category, t denotes term,n denotes thenumber of categories, $P(C_i|t)$ and $P(C_i|\bar{t})$ denotes conditionalprobabilities of category for the term appears in document andthe term does not appear in the document respectively.

Quick Reduct Algorithm

Quick Reduct Algorithm is an efficient algorithm for finding reduct. This is widely used is several soft computing implementations using Rough Sets. Quick Reduct algorithm proposed by A.Chouchoulas and Q. Shen. Quick Reduct algorithm attempts to calculate a reduct without exhaustively generating all possible subsets. It starts with an empty set and adds in turn, one at a time, those attributes that result in the greatest increase in the rough setdependency metric, until this produces its maximum possible value for the dataset. Information system is the main concept in rough set. With the help of approximation, rough set can defined. Here, there are some elements in U, where U is database. It has subset S. $R \subseteq U \times U$ is an equivalence relation. However, there is less knowledge is available about elements of U. R approximation is used for classification of set S. Lower approximation and upper approximation are sets of objects which decides object is surely classified as elements of set S or may classify as elements of set S respectively. Lower approximation and upper approximation are also known as positive region and negative region respectively and there difference is nothing but the boundary region of S. For rough boundary region a set 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. Only most informative feature is kept. Dependency measure in quick reduct algorithm is calculated using following formula given in equation 5:

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

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

Where P is a set of conditional attributes and Q is the decision attribute, $\gamma_p(Q)$ is the dependency between conditional attribute and decision attribute. If$k = 1$, Q depends totally onP; if$0 < k < 1$, Q depends partially on P, and if k=0 then Qdoes 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 as shown below [2]:
C: the set of all conditional attributes,
D: the set of decision attributes
QuickReduct (C, D)

1. Algorithm is starts with empty set $R - \{\}$
2. It adds one attribute at a time $T \leftarrow R$
3. For everyattribute, dependency is calculated with respect to decision attribute and then compared

$$\gamma_{RU\{x\}}(D) > \gamma_T(D)$$

4. If it provides the greatest increase in dependency metric then it gets add in the reduced set T ←R U{x}
5. 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 Quick Reduct algorithm is that it does not need any additional information about data for finding minimal set of attributes.

Naive Bayes Classifier

Naive Bayes is a simple technique for constructing classifiers models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. An advantage of Naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification.Conditionalprobabilities for each term against predefined category calculated by classifier only optimalcategory is assigned to web pages based on maximum posterior probability by considering the term frequencies. Hence, this methodis known as probabilistic method. Each training dataset can incrementally increase and decrease the probability of hypothesis. To get correct hypothesis nature of method should be incremental. In classification of web pages hypothesis means whether web page belongs to category or not.

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

Above formula is use for assigning web page to the specific category. Where c denote category, n denotes number of words in document.Laplacian correction is required for calculating conditional probability $P(word_1|c)$. The reason behind using laplacian correction is to prevent zero probabilities for terms which are not present in training examples. Formula is given for $P(word_1|c)$ using laplacian correction as follows in equation 7:

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

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, $P(c_1|document)$, $P(c_2|document)$,….., $P(c_m|document)$is compared. Document is assigned to category based on maximum value of $P(c\ \ |document)$calculated above which nothing but maximum posterior probabilityas shown in equation 8.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

$$\text{ArgMax}c_j \in c P(c_j|d_i) \approx \text{ArgMax}c_j \in c \prod_{k=1}^{n} P(tik|c_j)P(c_j) \quad (8)$$

IV. SYSTEM ARCHITECTURE

Block diagram of proposed architecture is as shown in Figure 1. It includes three parts. First part includes pre-processing, second part contains information gain method as feature selection method, rough set based quick reduct algorithm as dimensionality reduction method, and finally, third part is Naïve Bayesian classifier.

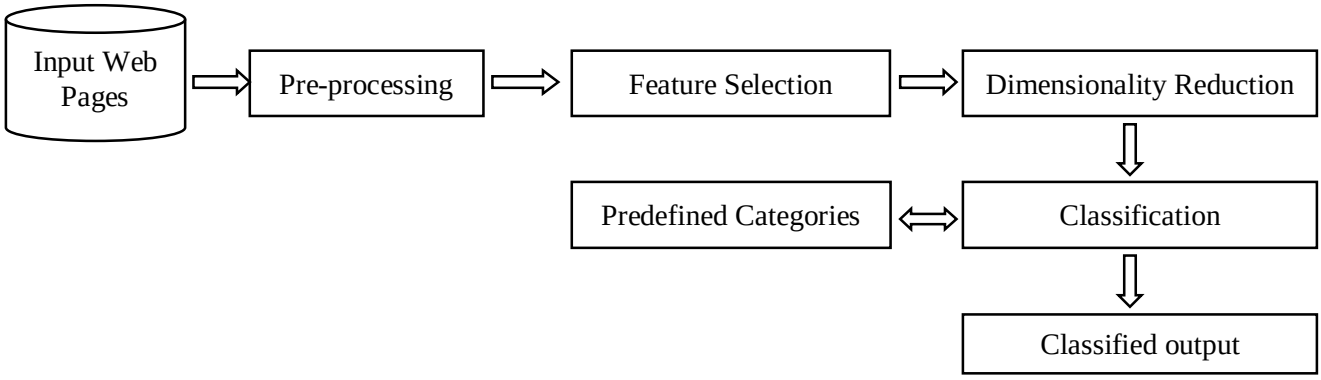

Figure 1: Architecture of feature selection and dimensionality reduction for web search result classification

a. PREPROCESSING

Pre-processingincludes three processes that are tokenization, stemming and stop word removal processes. In the process of tokenization file are tokenized into individual tokens using space as the delimiter. Stemming process converts words into their root/original form. For this process, Porter's stemming algorithm is used. Stop words are those, which do not convey any meaning and would be removed.

b. FEATURE SELECTION AND DIMENSIONALITY REDUCTION

After preprocessing of search results the words which are remain will be given to feature selection method. For each and every word, information gain will be calculated. In this calculation, entropy for each category is considered with probability of word and conditional probabilities of category against words. This calculationsare done by using formulae as shown in equation 2,3 and 4.$P(C_i|t)$and$P(C_i|\bar{t})$is the conditional probability of category against term if the term does not appear in the document and if the term appears in the document. In these equations, summation indicates that for every word all categories are considered. In the end of information gain, those words will be selected which are having value of information gain greater than or equal to a predefined threshold.

Next is Quick Reduct algorithm, Information system is required to apply the algorithm. Information system contains words as conditional attributes and category of web page as decision attributes. Using information system, correlation measure for each word will be calculated. Words with zero correlation measure are discarded as redundant terms. This indicates that remaining words provide the same quality of prediction as the original attributes.

c. NAÏVE BAYESIAN CLASSIFIER

For the classification of documents,classifier uses Conditional probabilities and Posterior probabilities. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.$P(C \mid document)$ is the conditional probability, calculated on the basis of probabilities of categories and

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

conditional probabilities of words against predefined categories as shown in equation 6. For conditional probabilities of words against predefined categories laplacian correction is used as shown in equation 7 to prevent zero probabilities for terms which are not present in training examples. Web pages are assigned to the most optimal category using maximum posterior probability value which is calculated using formula given in equation 8.

V. CONCLUSION

In this survey paper, the different efficient techniques for webpage classification are discussed. Techniques require more orless data for training as well as less computational time of thesetechniques.It fastens the access of web document searchingover the web to save the time and complexity of web pages.Rather than giving all words to classifier, only informative and relevant words are given to the classifier. Less informative and redundant terms removed using feature selection and dimensionality reduction methods.

REFERENCES

[1] Bo Tang, Haibo He, Paul M. Baggenstoss and Steven Kay, "A Bayesian Classification Approach Using Class Specific Features for Text Categorization", IEEE Trans. Knowl.Data Eng., vol.28, No. 6, June2016.
[2] ShraddhaSarode, JayantGadge, "Hybrid Dimensionality Reduction Approach for Web Page Classification".2015 ICCICT, Jan. 16-17, Mumbai, India.
[3] Juan Zhang, Yi Niu, HuabeiNie, "Web Document Classification Based on Fuzzy k-NN Algorithm", International Conference on Computational Intelligence and Security, IEEE, 2009.
[4] 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.
[5] Xiaoyue Wang, Zhen Hua, RujiangBai. "A Hybrid Text Classification model based on Rough Sets and Genetic Algorithms" Ninth ACIS International Conference on Software Engineering, Artificial
[6] Intelligence, Networking, and Parallel/Distributed Computing, IEEE, 2012.
[7] G.S. Tomar, ShekharVerma, AshishJha, "Web Page Classification using Modified Naïve Bayesian Approach", IEEE, 2006.
[8] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in Proc. Workshop Learn. for Text Categorization, 1998, vol. 752, pp. 41–48.
[9] Sneha V. Dehankar, K. P. Wagh, "Web Page Classification Using Apriori Algorithm and Naive Bayes Classifier" IJARCSMS volume 3, issue 4, April 2015, pg-527-533.
[10] Hetal C. Chaudhari, K. P.Wagh and P.N.Chatur," Search Results Clustering using TF-IDF Based Apriori Approach:A Survey Paper" IJECS Volume 4 Issue 1 January, 2015 Page No.10175-10179.
[11] Sneha K.Dehankar, K. P.Wagh and P.N.Chatur," Different Methods of Classification of Documents A Survey Paper", IJECS Volume 4 Issue 1 January, 2015 Page No.10180-10183.
[12] Sanjay D. Sawaitul, Prof. K. P. Wagh, Dr. P. N. Chatur," Classification and Prediction of Future Weather by usingBack Propagation Algorithm-An Approach",IJETAE,ISSN 2250-2459, Volume 2, Issue 1, January 2012.
[13] P. H. Govardhan, K. P. Wagh, P. N. Chatur," Web Document Clustering using Proposed Similarity Measure", International Journal of Computer Applications (0975 – 8887) National Conference on Emerging Trends in Computer Technology (NCETCT-2014).
[14] Pankaj G Devikar, Prof. Pankaj Sahu," The Use Of Artificial Neural Network ForCategorization And Indication Of Weather ForecastingDataset With Dynamic Library" IJERGS Volume 3, Issue 5, September-October, 2015 ISSN 2091-2730.
[15] Vishnu Kumar Goyal," A Comparative Study of Classification Methods in Data Miningusing RapidMiner Studio", (IJIRSE) International Journal of Innovative Research in Science & Engineering.