ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Classification of Web Search Results Using Modified Naïve Bayesian Approach

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

P.G. Student, Department of Computer Engineering, Govt. College of Engineering, Amravati, India[1]

Associate Professor, Department of Information Technology, Govt. College of Engineering, Amravati, India[2]

Head, Department of computer Science and Eng., Govt. College of Engineering, Amravati, India[3]

ABSTRACT:The primary purpose of data mining is to extract information from huge amounts of raw data. To get the useful data from large amount of available data is necessary. Web document classification includes the classification of web snippets into different categories based on their content. The classes are predefined in which the pages are classified. The web snippets from first three pages of Google extracted and prepossessed. Preprocessing includes tokenisation, reduction of redundant and irrelevant data. After the prepossessing of the web snippets, Modified Naïve Bayesian approach is used to get the snippets classified into predefined categories. From these the probability of each word will be calculated and page will be classified into its predefined class based on the highest posterior probability calculated. The Modified Naive Bayes classifier is used to calculate the probability of each word with respect to each class. By using snippets as a input we managed to reduce the require classification time up to 49.04 %, shows the F-measure value 93.79 % and achieved accuracy up to 96.01 %. An analysis of the system reveals that the snippets classification system works well even when the number of snippets is increased.

KEYWORDS: Modified Naïve Bayesian Classifier, Quick Reduct Algorithm, Tokenization, F-measure.

I. INTRODUCTION

Classification is the process of dividing the data into numbers of groups which are either dependent or independent of each other and each group act as a class. The task of classification can be done by several methods by different types of classifier. The main purpose of this work is to analyse the task of classification of web search results (snippets) into multiple classes and to learn how to achieve high classification accuracy in classifying the web search results (Snippets). The Text processing plays an important role in information retrieval, data mining, and web search. Text mining attempts to discover new previously unknown information by applying techniques from data mining. The user gets many results on web after submitting the query. Most of the links extracted are of not useful. The topic irrelevant data are shown for the required query, so to get the relevant data related to the keyword submitted, the classification of the data, documents are needed. In this the user will be getting the relevant data within less amount of time. The classes in which want to classify the extracted results are defined first and then the results extracted from the web are classified into the desired class based on its content. The probability for each word in the document is calculated and based on that probability; the document is classified into the desired class. The pre-processing of the web page is done first then after removing the stop words. From this then the probability for each word of document is calculated and document is then classified into the class it belongs based on the highest probability calculated. For this classification Nave Bayes classifier is used. The classification is done into three different categories.

II. RELATED WORK

ShraddhaSarode and JayantGadge[1] proposed a hybrid approach of dimensionality reduction for web page classification using a rough set and information 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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

and dimensionality reduction methods. From experimental results, we find that more than 60% less informative words are removed. To reduce the dimensionality of web pages Feature selection and dimensionality reduction methods are used. Information gain method and Rough set based Quick Reduct algorithm used for feature selection and for dimensionality reduction respectively. Web pages are classified using Naïve Bayesian method.

Sneha V. Dehankar, K. P. Wagh[2] 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 system because the value of true positive parameter for each keyword submitted is better while the value of false positive is less. The number of predefined class label and number of words that are predefine for each class can be increase for the better accuracy in future.

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 off with an empty set and adds in turn, one at a time, those attributes that result in the greatest increase in the rough set dependency metric, until this produces its maximum possible value for the dataset [3].

III. PROPOSED SYSTEM

A. Snippets as Input

The full text documents are not alwaysavailable due to several reasons, e.g., lack of access to a particular publication repository, invalid URL, files in different formats, *etc*. Secondly, snippets are usually short. In fact, we have to heavily rely on the title of each snippet, as it is often that there are no descriptions in a snippet. Thirdly, the quality of snippets is far from perfect. Due to the automatic extracted nature, snippets may contains errors or be incomplete, and descriptions of snippets could be empty or meaningless. Fourthly, domain knowledge is usually required even by humans to assess the similarity between two publications.

B. 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 attempts to calculate a reduct without exhaustively generating all possible subsets. It starts off with an empty set and adds in turn, one at a time, those attributes that result in the greatest increase in the rough set dependency metric, until this produces its maximum possible value for the dataset.

QUICK REDUCT (C , D)
Input: C - the set of all conditional features;
D - The set of decision features
Output: R- the feature subset

1. $R \leftarrow \{\}$
2. while $\gamma_R(D) \neq \gamma_C(D)$
3. $T \leftarrow R$
4. for each $x \in (C - R)$
5. if $\gamma_{R \cup \{x\}}(D) > \gamma_T(D)$
6. $T \leftarrow R \cup \{x\}$
7. $R \leftarrow T$
8. return R

C. Probability Calculation by Multinomial Naive Bayes Classifier

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Multinomial Naive Bayes is a specialized version of Naive Bayes that is designed more for text documents. Whereas simple naive Bayes would model a document as the presence and absence of particular words, multinomial naive Bayes explicitly models the word counts and adjusts the underlying calculations to deal with in.

Let us now discuss how multinomial naive Bayes computes class probabilities for a given document. Let the set of classes be denoted by C. Let N be the size of our vocabulary. Then MNB assigns a test document t_i to the class that has the highest probability $Pr(c|t_i)$, which, using Bayes' rule, is given by:

$$\Pr(c|t_i) = \frac{\Pr(c)\,\Pr(t_i|c)}{\Pr(t_i)}, c \in C \quad (1)$$

The class prior Pr(c) can be estimated by dividing the number of documents belonging to class c by the total number of documents. $Pr(t_i|c)$ is the probability of obtaining a document like t_i in class c and is calculated as:

$$Pr(t_i|c) = (\sum_n f_{ni})! \prod_n \frac{Pr(w_n|c)^{f_{ni}}}{f_{ni}!} \quad (2)$$

Where f_{ni} is the count of word n in our test document t_i and $Pr(w_n|c)$ the probability of word n given class c. The latter probability is estimated from the training documents as:

$$\widehat{Pr}(w_n|c) = \frac{1 + f_{nc}}{N + \sum_{x=1}^{N} f_{xc}} \quad (3)$$

where F_{xc} is the count of word x in all the training documents belonging to class c, and the Laplace estimator is used to prime each word's count with one to avoid the zero-frequency problem. The normalization factor $Pr(t_i)$ in Equation can be computed using

$$Pr(t_i) = \sum_{k=1}^{|C|} Pr(k) Pr(t_i|k) \quad (4)$$

Note that that the computationally expensive terms $(\sum_n f_{ni})!$ and $\prod_n f_{ni}!$ in Equation 5 can be deleted without any change in the results, because neither depends on the class c, and Equation 5 can be written as:

$$Pr(t_i|c) = \alpha \prod_n Pr(w_n|c)^{f_{ni}} \quad (5)$$

Whereα is a constant that drops out because of the normalization step.

D. Mathematical Parameters for Comparison

1. **Recall**: Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labelled as belonging to the positive class but should have been).

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

2. **Precision**: The precision for a class is the number of true positives (i.e. the number of items correctly labelled as belonging to the positive class) divided by the total number of elements labelled as belonging to the positive class *i.e.*(the sum of true positives and false positives, which are items incorrectly labelled as belonging to the class).

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

Where, True Positive (TP) refers to the number of documents correctly classified to that category, False Positive (FP) refers to the number of documents incorrectly rejected from that category, True Negative (TN) refers to the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

number of documents correctly rejected from that category, False Negative (FN) refers to the number of documents incorrectly classified to that category.

3. **F-measure**: A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

4. **Accuracy**: The term accuracy in general for evaluating systems or methods refers to the bias of predictions, i.e. it answers the question how good predictions are on average.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

E. Example for Classification

In this example we will present the mathematical evaluation of results and will do the analysis by considering the precision, recall. F-measure (F1 Score) and Accuracy followed by their individual average. For the analysis of system we take three search queries such as "micromax, superman and tesla" on which we first perform Web Page classification followed by Snippet Classification.

IV. MATHEMATICAL ANALYSIS

This section gives brief idea about comparison of computational analysis and experimental analysis. Here we are compare both model on the basis of some analysis parameter. We calculated all the parameter for comparison of both the system. Average value of all parameters for all three queries is as follows:

Let,

WPC → Web Page Classification (Existing System)

SC → Snippets Classification (Proposed System)

Table 1: Average Recall Value

	WPC (%)	SC (%)
Micromax	69.16	83.33
Superman	96.49	100
Tesla	94.44	100

Table 2: Average Precision Value

	WPC (%)	SC (%)
Micromax	85.18	86.74
Superman	77.77	100
Tesla	90	100

Table 3: Average F-measure Value

	WPC(%)	SC(%)
Micromax	65.84	81.39
Superman	81.48	100
Tesla	91.09	100

Table 4: Average Accuracy Value

	WPC(%)	SC(%)
Micromax	82.55	88.05
Superman	93.64	100
Tesla	93.69	100

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

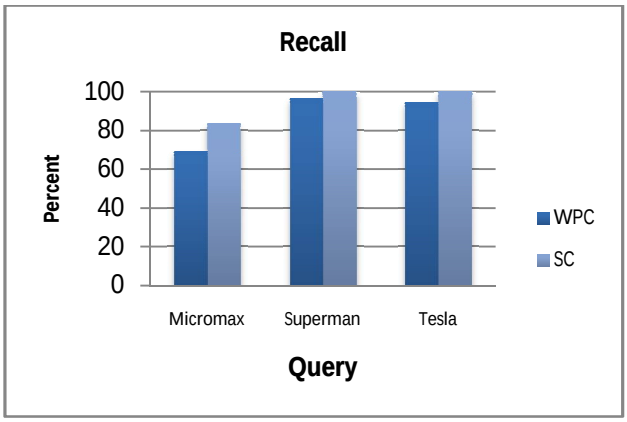

Fig 1: Recall Value Comparison

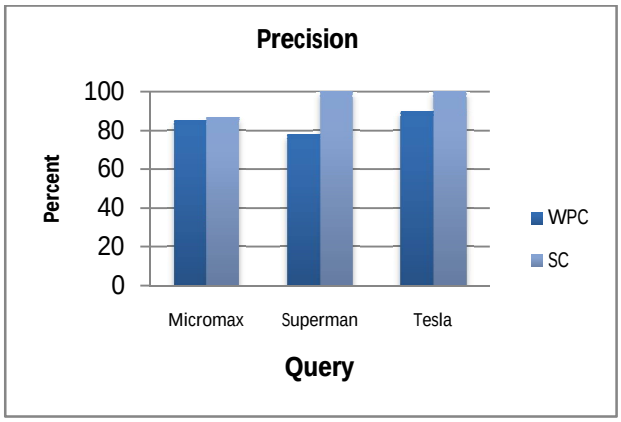

Fig 2: Precision Value Comparison

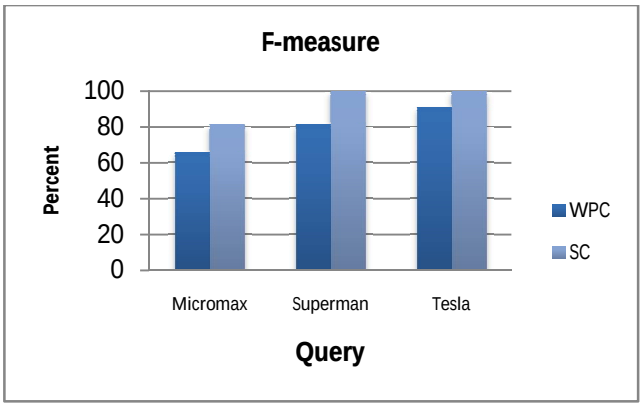

Fig 3:F-measure Value Comparison

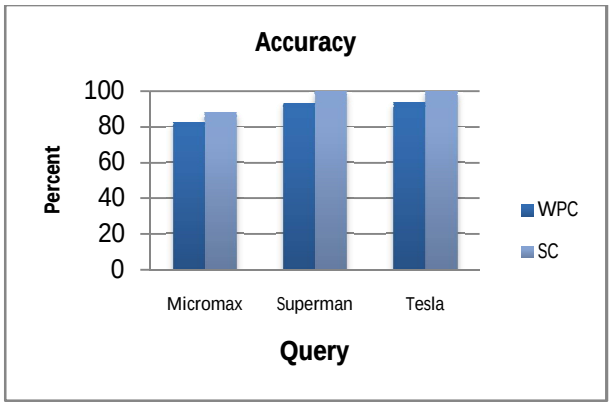

Fig 4:Accuracy Value Comparison

Table 5: Analysis of required Time

Query	Time for WPC (ms)	Time for SC (ms)
Micromax	1311.96	891.935
Superman	2110.62	956.952
Tesla	2185.82	865.091
Avg. Time	1869.46	904.65

Table 6: Comparison between Average Mathematical Parameter

	WPC (%)	SC (%)
Recall	89.96	96.01
Precision	86.69	94.44
F-1	84.31	95.58
Accuracy	79.47	93.79

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

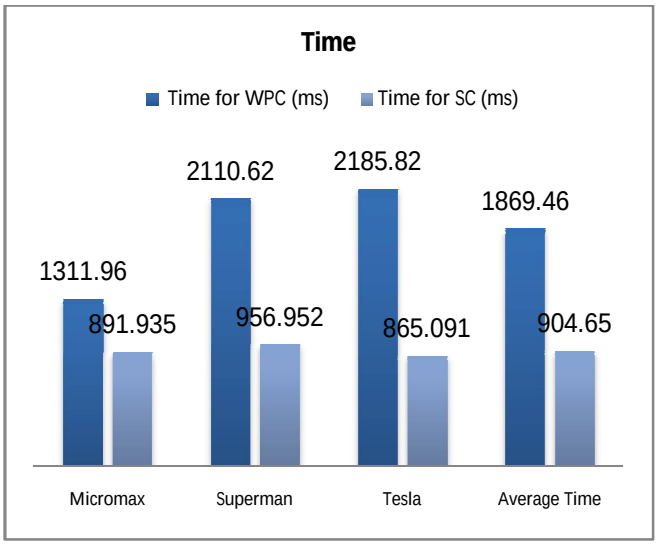

Fig 5: Analysis of time

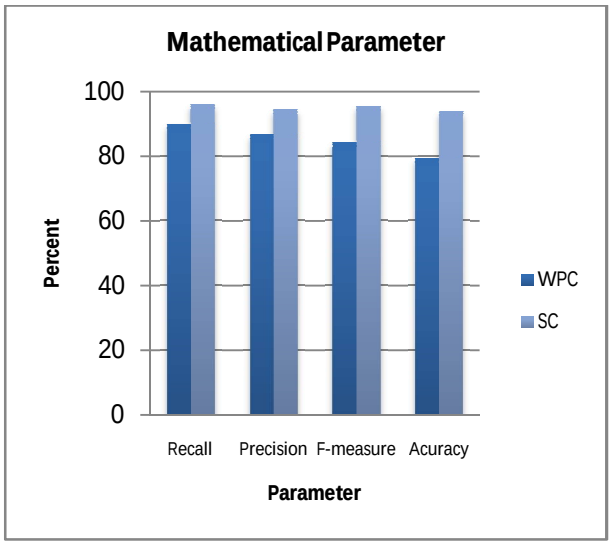

Fig 6: Analysis of Mathematical Parameters

All the table above shows the average recall, precision, f-measure and accuracy respectively, from this table we clearly depict that our proposed system is better than existing system. The figures show the graphical representation of all the mathematical parameters along with time required for completion of classification.

V. CONCLUSIONS

Snippets Classification system classifies the number of web snippets that are taken as input into the multiple classes. Snippets we used, acts as a light weight input. Redundant, irrelevant words are removed from every extracted snippet after preprocessing. Which makes snippets more informative for classification, Modified Naive Bayesian Approach for classification gives better accuracy for classifying the web snippets on the basis of their content.

Web snippets are classified on the basis of highest probability calculated for each of the web snippets. Classifying the snippets on the basis of their content gives the better accuracy rather than focusing on the HTML and URL tags of snippets for classification. By using snippets as a input we managed to reduce the require classification time up to 49.04 %, shows the F-measure value 93.79% and achieved accuracy up to 96.01 %. All the inputs are classified and displays according to their category label which shows the improved data availability and fasten data access, As a result information retrieval and accuracy in content delivery on the web are improved, which helps to increase user's interest.

REFERENCES

1. ShraddhaSarode, JayantGadge, "Hybrid Dimensionality Reduction Approach for Web Page Classification", 2015 International Conference on Communication, Information Computing Technology (ICCICT), Jan 16-17, Mumbai, India.
2. 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.
3. K.Anitha , Dr.P.Venkatesan,"Feature Selection By Rough Quick Reduct Algorithm",International Journal of Innovative Research in Science, Engineering and Technology (ISO 3297: 2007 Certified Organization) Vol. 2, Issue 8, August 2013
4. Janos Abonyi, BalazsFeil, "Computational Intelligence in Data Mining", Informatica 29 (2005) 3–12.
5. Xiaoguang Qi, Brian D. Davison, "Web Page Classification: Features and Algorithms", ACM Computing Surveys, Vol. 41, No. 2, Article 12, Publication date: February 2009.