RESEARCHERID
THOMSON REUTERS

[Ali, 10(1): Jan.-Mar. - 2020] ISSN 2277 – 5528
Impact Factor- 5.085

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & MANAGEMENT

"AN ENHANCED DOMAIN CLASSIFICATION OF RANDOM OFFLINE AND ONLINE DATA"

Rashid Ali[1], Dr.Dhanraj Verma[2]
Dept. Of Computer Science & Engineering, Dr.APJ Abdul Kalam University, Indore, India.

ABSTRACT

The search engine like Google provides record of results that shows a list of ranked output. The ranking does not consider the subject of the file. The results of search engine are not in a well- defined group. This may also be frustrating, as the users have to scroll through many inappropriate results. This could come up when the user is a beginner or has superficial capabilities about the domain of interest, however more as a rule it is due to the question being brief and ambiguous.

One answer is to organize search results through categorization, in specific, the classification. A goal of testing is to test designed on a controlled data set, which shows that classification- bounded search could enhance the person's search expertise in terms of the numbers of results the person would must check out earlier than pleasing his/her query.

This work uses the naive bayes classifier, which is a simple and effective method for establishing classifiers. The proposed model for finding domain, related to user query based on document index matrix. The proposed implementation combine the both approach simultaneously which is term based and phrased based. Document index matrix used term, phrased based document matrix in such a manner that it is compare with training data, and put them into relatively domain. The naïve bayes algorithm used to find maximum probability occurrence from both the matrix. The output comes in the form of suggestion domains list. Users easily retrieve the data with minimum time. An experimental result shows the proposed work is better than previous work.

Keywords: Data Mining, Data sets, Association rule mining, naive bayes classifier, domains list, frequent item set, large database.

I. INTRODUCTION

The use of digital text increases as the social media increases their effect in daily life. A number of research groups and individual researchers are working to finding the patterns on these data. This study represents the search engine optimization analysis and sentiment analysis techniques with a new classification algorithm for enhancing the performance of text classification. The given chapter provides an overview of the proposed work and involved investigation.

Knowledge mining is a procedure of mining information from the raw information. Extraction of the similar text from a raw set of text is the generation of text data mining. Clearly, text data belongs to an unstructured method and labelling of information is tricky undertaking, therefore, many of the applications are utilizing the classification approaches for categorizing knowledge. Text classification captures the relevant result for each user query, Naïve Bayes classifier is a simple and effective approach to classify text document, which uses probabilistic classification technique. Naïve Bayes classifier using Bayes' theorem for classifying unknown retrieved data from Google search engine and few modifications are there to increase performance of classifier.

RESEARCH OBJECTIVE

The offered study is inquisitive about discovering probably the most useful techniques for textual content evaluation and categorization. Some important keyword phrases are concerned with the present objective.

1. Text analysis method Detection: This phase uses the strong literature for finding the most appropriate classification technique that promises to provide the text pattern analysis for the specified attributes from both the class distributions.
2. Design and implementation of classifier: Different articles and research on the text classification techniques are studied, this work uses the Naive Bayes classifier for text classification, and few modifications applied to increase performance.
3. Analysis of proposed technique: Analysis of proposed technique section shows the developed classifier that verifies on a series of actual web based world knowledge. Additionally the efficiency of the method depends on accuracy, time, and area complexity

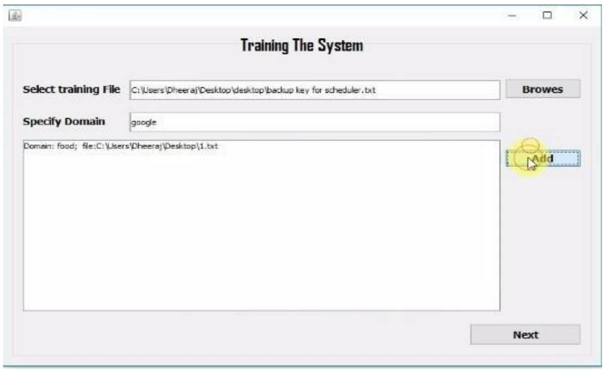
II. TECHNICAL APPROACH AND ANALYSIS

1. Methodology of the Study

Different UML diagrams (such as use-case diagram, system context diagram and state chart diagram) describe the desired software (proposed classifier) features in detail. These diagrams are intended to explain the software in satisfactory detail that programmers may build up the software with minimum extra effort.

2. Working procedure

The main working process of a proposed implementation as follows: First, select the categories for training.



Select training File: C:\Users\Dheeraj\Desktop\desktop\backup key for scheduler.txt

Specify Domain: google

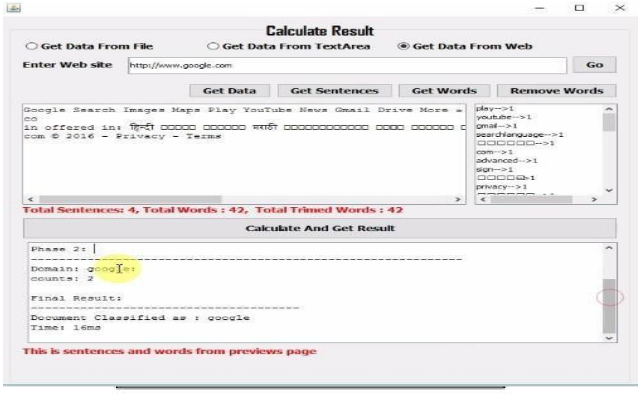
Domain: food; file:C:\Users\Dheeraj\Desktop\1.txt

Figure: 01 Training the System

This module used to train the system for different domains. Special keyword, which is associated with particular domain, is used for training.

After the training procedure, when we give input as a query the related categories found.

The testing procedure of a particular domain finder. It shows the various process involve in this mechanism. When user gives any input document.



Enter Web site: http://www.google.com

```
Google Search Images Maps Play YouTube News Gmail Drive More »
co
in offered in: हिन्दी □□□□□ □□□□□□ मराठी □□□□□□□□□□□□ □□□□ □□□□□□ □
com © 2016 - Privacy - Terms
```

```
play-->1
youtube-->1
gmail-->1
searchlanguage-->1
□□□□□□-->1
com-->1
advanced-->1
sign-->1
□□□□□□>1
privacy-->1
```

Total Sentences: 4, Total Words : 42, Total Trimed Words : 42

```
Phase 2:
------------------------------------------------------------------
Domain: google:
counts: 2

Final Result:
---------------------------------------
Document Classified as : google
Time: 16ms
```

This is sentences and words from previews page

Figure: 02 Calculate Result

Firstly it will create the sentences on it and then it will fit into term matrix and phrased matrix.

III.TEST RESULTS AND DISCUSSIONS

Implementation of the proposed system gives results of the proposed classification technique and previously available technique and compares using their performance graphs. The given chapter provides the detailed discussion about the preformed experiments and their results.

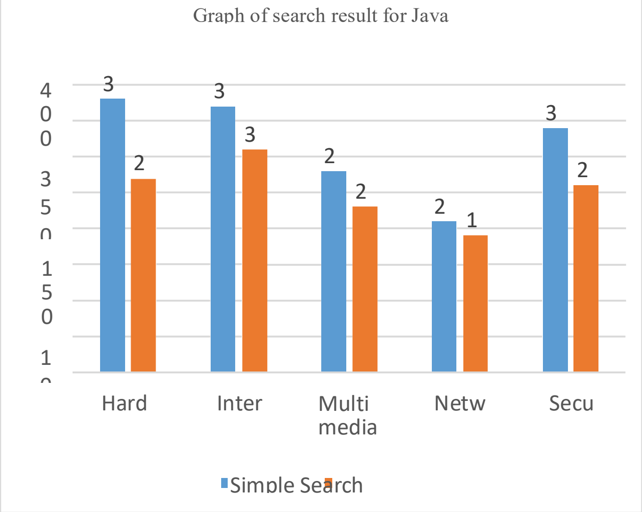
Accuracy: The amount of correctly recognized patterns is as the classification accuracy, in a data mining based classification system. The accuracy of the system in terms of percentage, computed using the following formula:

Accuracy = (accurately classified patterns / total input patterns) X 100 (5.1)

Time consumption: The amount of time required to classify the entire test data, known as the time consumption. Following formula computes the time consumption:

Time consumed =end time-start time (5.2)

Result analysis gives the brief description about the comparative time consumption of the proposed and tradition algorithms, which shows proposed algorithm performs well as compare to traditional algorithm.



Graph of search result for Java

We calculate the result related to its domain. When user gives the query, proposed implementation search by comparison the term and phrased used in this document. Related domain find after followed some procedure.

IV.CONCLUSION

Domain name classification is a great challenge in a web mining. Thousands of domains are there, find the right one from them is quite difficult process. Searching process takes too much time and many filtrations to search out the right one from them. Naive Bayes classifier comes to solve the problem of domain classification. Proposed work uses term and phrased based approach simultaneously to get the accurate result from the training domain. A new query result returns document, when that document enters, Naive Bayes uses to find the probability of highest term and phrased available there using matrix. A modified Naive Bayes algorithm exists to deal with that filter result. Modified Naive Bayes works on selected resulted whose frequency of occurrence is high, so the modified Naive Bayes performance gets higher in terms of timing. Experimental results show the presented work outperforms far better than the previous one.

REFERANCES

[1] Samuel Ieong, Nina Mishra, Eldar Sadikov, Li Zhang, "Domain Bias in Web Search" , ACM New York, NY, USA ©2012.

[2] Bo Geng, Linjun Yang, Chao Xu, Xian-Sheng Hua, "Ranking Model Adaptation for Domain- Specific Search", IEEE Transactions on Systemsknowledge and data engineering vol. 24, no. 4, April 2012.

[3] Junghoon Chae, Dennis Thom, Yun Jang, Sung Ye Kim, Thomas Ertl, David S. Ebert, "Public behaviour response analysis in disaster events utilizing visual analytics of microblog data",Elsevier Ltd. All rights reserved 2013.

[4] Umajancy. S, Dr. Antony Selvadoss Thanamani, "An Analysis on Text Mining –Text Retrieval and Text Extraction", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, August 2013.

[5] Xia Hu, Lei Tang, Jiliang Tang, Huan Liu, "Exploiting Social Relations for Sentiment Analysis in Microblogging", WSDM '13, February 4–8, 2013, Rome, Italy, ACM 978-1-4503- 1869-3/02/ 2013.

[6] Rahul A. Patil, Prashant G. Ahire, Pramod. D. Patil, Avinash L. Golande, "A Modified Approach to Construct Decision Tree in Data Mining Classification", International Journal of Engineering and Innovative Technology (IJEIT), Volume 2, Issue 1, July 2012

[7] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. ACM SIGMOD, 1993.

[8] Vishal Gupta, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009