Volume 5, Issue 5, MAY 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

www.ijarcsse.com

Research Paper

Available online at: www.ijarcsse.com

Automated Text Categorization of Advertisemnets

Mansi Naik, Falguni Ahir, Priyanka Puvar, Bhagirath Prajapati
A.D. Patel Institute of Technology
Gujarat, India

Abstract*—Text categorization has become most active research topic in the field of information retrieval. The main objective of text categorization is to assign the entries from a large set of pre-specified categories to a document. Categories may be derived from a sparse classification scheme or from a large collection of very specific identifiers. Traditionally the task of categorization is performed by domain experts. First the incoming document read and comprehended by the expert and then assigned a number of categories chosen from the set of pre-defined categories. It is inevitable that a large amount of manual is required. To deal with this kind of problem is to learn a categorization scheme automatically from training example. Once it is learned it can be used for classify future documents. It involves issues commonly found in machine learning problems. This approach is typically construct a classifier for each category and the categorization process becomes a binary decision problem for the particular category. One useful application for automatic categorization is to support effective text retrieval. Apart from studying the effectiveness of automatic categorization directly, the second objective of this report is to investigate application of this categorization process to text retrieval. By introducing classifiers which use a base model like bayesian independence classifiers for categorization it can reduce the large amount of manual effort. It also helps to investigate the application of this categorization process to text retrieval.*

Keywords*— Text Classification, Machine Learning Techniques, Data Representation*

I. INTRODUCTION

Text classification has recently become an active research topic in the area of information retrieval. The main objective of text categorization is to assign entries from a set of pre-specified categories to a document.[1] A piece of paragraph or text is defined here as a document. Simple Bayesian classifiers is popular for classification of the text , the probabilistic approaches make strong assumptions about how the data is generated, then they use a collected labeled training examples to estimate and rectify the parameters of the generative model. Categorization on new paragraph or a text is performed by selecting the class that is most likely to generate the best example The naive Bayes classifier is the simplest of all models, in which it assumes that all attributes of the examples are independent of each other given the context of the class. This is the so-called “naive Bayes assumption.” While this assumption is clearly false in most real-world tasks, naive Bayes often performs classification very well. Because of the independence assumption, the parameters for each attribute can be learned separately, and this greatly simplifies classification of text is just such a domain with a large number of attributes. The attributes of the examples to be classified are piece of text and the paragraph different words can be quite large indeed. The algorithm has been successfully applied to classification in many research efforts. [9]The second objective of this report is to investigate application of this categorization process to text retrieval. Using an automated text categorization of advertisement ,we can enhance the feature of an application by reducing manual work by domain experts. To reduce large amount of manual efforts by introducing classifiers which use a base model like bayesian independence classifiers for categorization.

II. PROBLEM STATEMENT

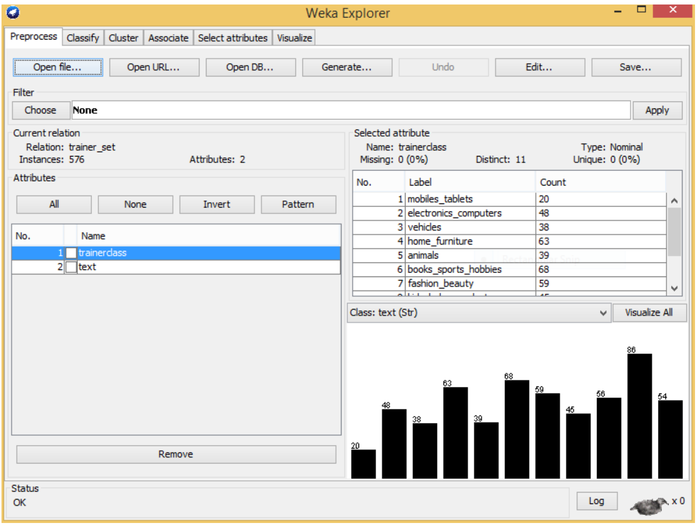
In the first phase of text classification, raw data is needed on which further classificaton is being processed. So, Data collection is the process of gathering information- a general word which is used in the system is called **Raw Data.** The dataset collection is common to all fields of study including physical and social science, humanities, business, etc. Here [2] **D** is a entire set of **Documents**, where d_i is one of the document of entire set. Documents which is set of text or paragraph are taken from the OLX website which is the Raw Data set for the system . **{ c_1, c_2, ..., c_n}** is a set of all **Categories** that are also fetched from the home page of the OLX website, these categories will be considered as pre-specified for the system, then the categorization of text is done by one category c_j to a document d_i . Here the documenst which are text or set of paragraphs are fetched from the different World Wide Webs by creating the **Web Crawler** for the system. D is the set of documents that has been stored in a database for classification and further text retrieval.[10] **Training Set T** which is in the form **{ c_1 ,'d_i' \n c_2 ,'d_i' \n c_2 ,'d_i' \n .... c_n ,'d_i' }** is to be developed for the learning process of the system due to which the system gets trained and for further Testing process . Training file which is in the format of **.arrf** that will run in WEKA software. [3] **Weka** (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, will be used for training purpose of the system. [10]WEKA is used for graphical presentation of classifeid text.

III. PROPOSED WORK

[4]A Web crawler is an Internet bot that browses the World Wide Web, for Web indexing. A Web crawler is also known as **Web spider** , an automatic indexer or a Web scutter. Web search engines uses Web crawling or spidering software to update their web content or indexes of others sites' web content. [11]Web crawlers also has the feature of copying all the pages they visit for later processing by a search engine that keeps the record of the downloaded pages so that users can search and access them much more quickly. Crawlers are also used to validate hyperlinks and HTML code. They can also be used for web scraping to search and retrieve the required information or data**.** **jsoup**[5] is one of the library in java which is also called as Java HTML Parser, which works with the real-world HTML. It provides a very convenient API of extracting datas and manipulating those data, using the best of Cascading Style Sheet(CSS), Document Onject Model (DOM),etc. jsoup implements the WHATWG HTML5 specification, and parses HTML to the same Document Object Model as modern browsers do. [8]They scrape and parse HTML by their tags by fetching different text, strings etc .They help us to find and extract data traversal selectors . They use to manipulate the HTML elements and text .jsoup is used to prevent the XSS attacks.

IV. IMPLEMENTATION WORK

Firstly, data is collected which is called as a document in the form of the text or a paragraph from different websites via Socket Programming which is actually known as a network socket. The process is an endpoint of an inter-process communication flow which takes place across a computer network. In today’s world, most communication between computers is based on the Internet Protocol; therefore most network sockets are Internet sockets. Raw Data is stored in database. The system works on algorithm called j48 algorithm .J48 is an open source Java implementation of the C4.5 algorithm . It is used in Weka data mining tool. [6][7]C4.5 is a program that designs a decision tree based on a set of labeled input data. The decision trees generated by C4.5 can be used for classification of the raw data of the system. For training of the system, Weka tool is being used which gives the graphical presentation of the classified advertisements with respect to the pre-specified categories. The file which is to be run in Weka is of **.arff** extension . **Weka** (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, will be used for training purpose of the system. WEKA is used for graphical presentation of classifeid text. The file will be generated as .dat file by the weka file .arff extension. Once algorithm is trained as .dat file, it will classify the sample data which is here refer as raw data set.



Filter: None

Current relation
Relation: trainer_set
Instances: 576
Attributes: 2

Attributes

No.		Name
1	□	trainerclass
2	□	text

Selected attribute
Name: trainerclass
Type: Nominal
Missing: 0 (0%)
Distinct: 11
Unique: 0 (0%)

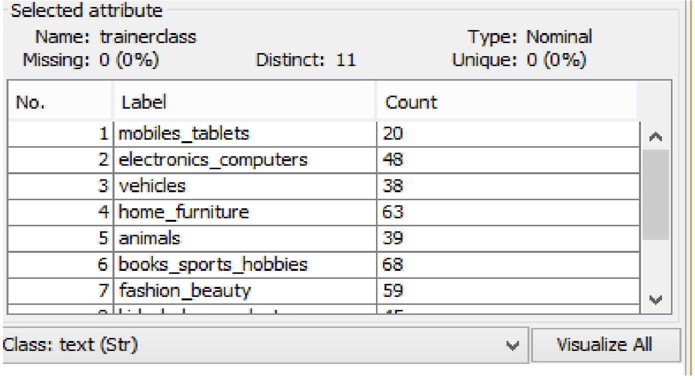
No.	Label	Count
1	mobiles_tablets	20
2	electronics_computers	48
3	vehicles	38
4	home_furniture	63
5	animals	39
6	books_sports_hobbies	68
7	fashion_beauty	59

Class: text (Str)

Status
OK

Fig:4.1

The training process is done in Weka tool that is displayed in fig 4.1that shows how the Weka toolkit works. Fig 4.2 depicts the names of the pre-specified categories that are listed in a list and in the other field count suggests how many entries are present in each categories.



Selected attribute
Name: trainerclass
Type: Nominal
Missing: 0 (0%)
Distinct: 11
Unique: 0 (0%)

No.	Label	Count
1	mobiles_tablets	20
2	electronics_computers	48
3	vehicles	38
4	home_furniture	63
5	animals	39
6	books_sports_hobbies	68
7	fashion_beauty	59

Class: text (Str)

Fig 4.2

Fig 4.3 displays the graphical presentation of the trained data in Weka toolkit.

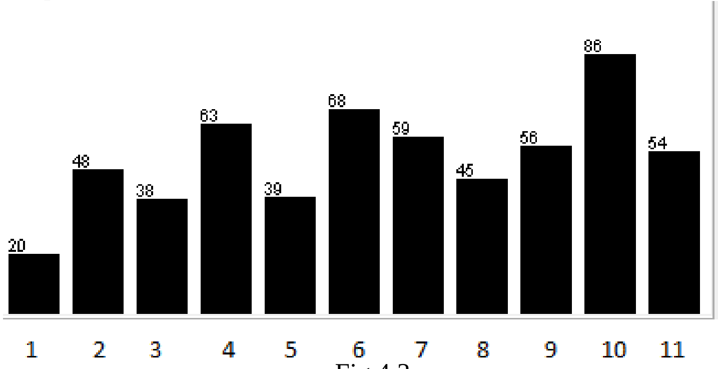
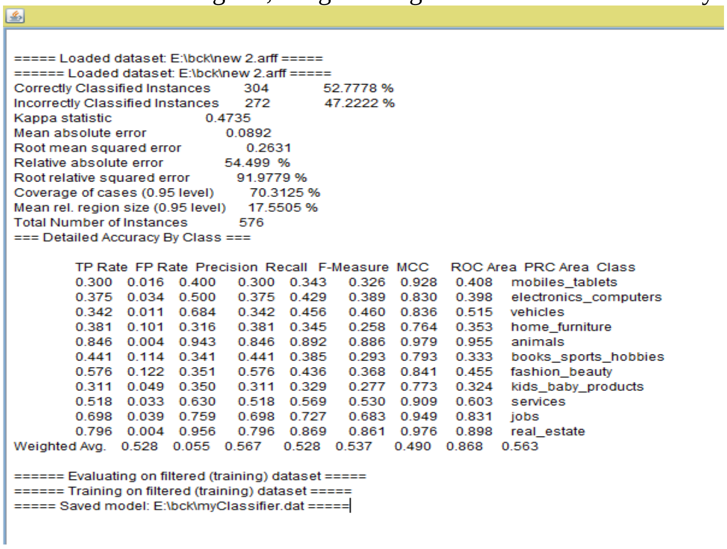


Fig 4.3

After training dataset is created, j48 classifer algorithm is used that helps to classify the data.It will generate .dat file for the trained data and then classification of fetched data will start and data is fetched from all the possible different online Advertisement websites which are OLX. .dat file is generated and is shown in fig:4.4 Where TP rate, FP rate, Precision, Recall, F-Measure, MCC, ROC Area, PRC Area of every categories of the system is calculated as well as error rates also calculated which includes Mean Absolute error, Root Mean Squared error, Relative Absolute error, Root Relative Squared error. Every other detail is shown in fig:4.4, weighted avg is also calculated for every categories of the system.



```
===== Loaded dataset: E:\bck\new 2.arff =====
====== Loaded dataset: E:\bck\new 2.arff =====
Correctly Classified Instances       304             52.7778 %
Incorrectly Classified Instances     272             47.2222 %
Kappa statistic                    0.4735
Mean absolute error                   0.0892
Root mean squared error                 0.2631
Relative absolute error              54.499  %
Root relative squared error            91.9779 %
Coverage of cases (0.95 level)          70.3125 %
Mean rel. region size (0.95 level)      17.5505 %
Total Number of Instances             576
=== Detailed Accuracy By Class ===

             TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
             0.300    0.016    0.400      0.300    0.343      0.326    0.928     0.408     mobiles_tablets
             0.375    0.034    0.500      0.375    0.429      0.389    0.830     0.398     electronics_computers
             0.342    0.011    0.684      0.342    0.456      0.460    0.836     0.515     vehicles
             0.381    0.101    0.316      0.381    0.345      0.258    0.764     0.353     home_furniture
             0.846    0.004    0.943      0.846    0.892      0.886    0.979     0.955     animals
             0.441    0.114    0.341      0.441    0.385      0.293    0.793     0.333     books_sports_hobbies
             0.576    0.122    0.351      0.576    0.436      0.368    0.841     0.455     fashion_beauty
             0.311    0.049    0.350      0.311    0.329      0.277    0.773     0.324     kids_baby_products
             0.518    0.033    0.630      0.518    0.569      0.530    0.909     0.603     services
             0.698    0.039    0.759      0.698    0.727      0.683    0.949     0.831     jobs
             0.796    0.004    0.956      0.796    0.869      0.861    0.976     0.898     real_estate
Weighted Avg.  0.528    0.055    0.567      0.528    0.537      0.490    0.868     0.563

====== Evaluating on filtered (training) dataset =====
====== Training on filtered (training) dataset =====
===== Saved model: E:\bck\myClassifier.dat =====
```

Fig 4.4

V. RESULT AND ANALYSIS

The text or a document is now classified. Here, the classification for the document is done for every cities of India, and graphical presentation is also shown. Here the fig 5.1 pie chart depicts the classification of the advertisements for pune where maximum advertisement have been posted for Fashion and beauty category. And least advertisements are posted for Electronics. It depicts that which city prefer which kind of product the most.

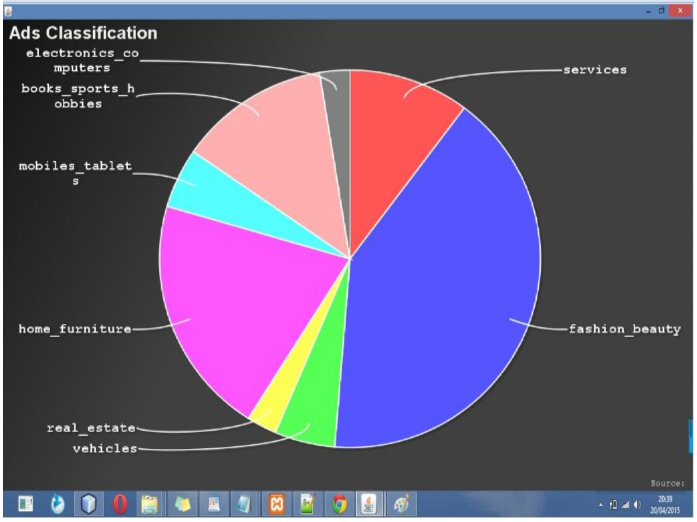


Fig 5.1

VI. CONCLUSION

The document which is a set of paragraphs that had been fetched from the olx website by creating web crawler and stored in a text file. Web Crawler is designed by using jsoup library . The training data set is created and trained using the WEKA toolkit which also shows the graphical representation of the trained data. Classification of document is performed by using j48 classifier algorithm. And finally pie chart of the classified data is obtained for every cities. Limitation of the system is some documents may fall into two categories due to some accuracy reasons, due to which training set has to be modified and updated at every interval of time due to which will get accurate results. Efficiency has to be obtained for some confusing text that belongs to one or more categories.

VII. FUTURE WORK

The system needs to be modifeied in the section in which more categories could be added as well as deleted if it is no more used. The efficiency has to be achieved by adding features to the pre-specified categories due to which no text falls into more than one category.

ACKNOWLEDGMENTS

This research paper was supported by the Project GuideProf. Bhagirath Prajapati and Prof. Priyanka Puvar as well as the institution A. D. Patel Institute of Technology, Gujarat, India.

REFERENCES

[1] A Tutorial on Automated Text Categorisation Fabrizio Sebastiani Istituto di Elaborazione dell'Informazione Consiglio Nazionale delle Ricerche Via S. Maria, 46 - 56126 Pisa (Italy) E-mail: fabrizio@iei.pi.cnr.it

[2] M. IKONOMAKIS , S. KOTSIANTIS, V. TAMPAKAS ,University of Patras, GREECE " Text Classification Using Machine Learning Techniques" August 2005,pp. 966-974

[3] WEKA_by http://en.wikipedia.org/wiki/Weka

[4] Web_Crawler_by::Wikipedia[http://en.wikipedia.org/wiki/web_crawler]

[5] jsoup by jsoup org[http://jsoup.org/]

[6] PERFORMANCE TUNING OF J48 ALGORITHM FOR PREDICTION OF SOIL FERTILITY Jay Gholap Dept. of Computer Engineering College of Engineering, Pune, Maharashtra, India

[7] C4.5 Machine learning – ACM Digital Library

[8] HTML Parser- http://htmlparser.sourceforge.net/

[9] Automated Text Categorization :Tools, Techniques and Applications ,Fabrizio Sebastiani Istituto di Elaborazione dell'Informazione Consiglio Nazionale delle Ricerche56124 Pisa Italy,Email:fabrizio@iei.pi.cnr.it, [WWW:http://faure.iei.pi.cnr.it/~fabrizio/c_]

[10] Training_set_of_WEKA https://www.youtube.com/watch?v=uiDFa7iY9yo

[11] Crawling pages [https://code.google.com/p/crawler4j/]

[12] Web Mining Research: A SurveyRaymond KosalaDepartment of Computer ScienceKatholieke Universiteit LeuvenCelestijnenlaan 200A, B3001Heverlee, BelgiumRaymond@cs.kuleuven.ac.beHendrik BlockeelDepartment of Computer ScienceKatholieke Universiteit LeuvenCelestijnenlaan 200A, B3001Heverlee, BelgiumHendrik.Blockeel@cs.kuleuven.ac.be