INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

APACHE TIKA: AN EFFECTIVE DATA MINING TOOL FOR MINING REGULATORY RULES.

SHRADDHA SURATKAR, AISHWARYA BHURKE, ANKITA BHUSARI, APURVA CHOUDHARY

Department of Computer Engineering, DYPIEMR, Akurdi, Pune – 33

Accepted Date: 15/03/2016; Published Date: 01/05/2016

Abstract: Regulatory compliance is a significant need in today's industry based world. Regulatory rules are the strategies provided to an organization for ensuring that the products manufactured by them are not biased. Regulatory bodies like NERC, FDA propose regulatory rules for various industries such as utility, banking, pharmaceutical, energy, food and insurance. These rules vary according to each area but are common to all industries related to that particular domain. Despoliation of these rules may lead to tremendous amount of losses in terms of economy and production. The industries are then likely to pay penalties in terms of money. They may be forced down to stop their production. The stock value and brand value is thus at high risk. There is direct business loss as well as high cost is involved in relaunch. Therefore industries need to extract the relevant rules accordingly from the regulatory documents. In order to extract relevant rules, they need a way of mining text as these rules have a very high impact on them. It is mandatory for the industries to keep a record of every updated rule and follow them. Every industry must have a separate team to look after these updates and inform the officials. The team should go through each and every rule that may or may not be of their interest. But this process becomes tedious and incurs high cost. To minimize the cost and simplify the work of industry, a system is proposed which classifies the regulatory rules that are applicable to that specific industry using a suitable data mining tool (Apache Mahout - Tika).

Keywords: Apache Mahout, Apache Tika, Classifier, Training dataset

Corresponding Author: MISS. SHRADDHA SURATKAR

Access Online On:

www.ijpret.com

How to Cite This Article:

Shraddha Suratkar, IJPRET, 2016; Volume 4 (9): 503-514

PAPER-QR CODE

1. INTRODUCTION

Industrial sectors have recorded a remarkable growth in recent years. Manufactured products from these industries are becoming an important part of our daily routine. These products effect on each class of people in diverse ways. There are various regulatory bodies who look after the manufactured products. It is their responsibility to make sure that these products are safe so that no adverse effects are caused on the consumers.

Regulatory bodies like NERC, FDA, thus, proposes regulatory rules for various industries such as pharmaceutical, energy, food, utility, banking and insurance. These rules vary according to each domain but are generic to all industries related to that particular domain. Violation of these rules may lead to tremendous amount of loss in terms of economy and production. The company is then liable to pay penalties in terms of money or may be forced down to stop their production. The stock value and brand value is thus at high risk. There is direct business loss as well as high cost is involved in relaunch. Therefore, Regulatory compliance is important. In order to extract the relevant rules accordingly from the regulatory documents, industries need a way of mining text. These rules have a very high impact on them. It is mandatory for the industries to keep a record of every updated rule and follow them. Every industry thus have a team to look after these updates and inform the officials. The team have to go through each and every rule that may or may not be of their interest. Text categorization is a technique often used as a basis for application of document processing and analysis. The process becomes very lengthy, tiresome and incurs high cost. To simplify their work and minimize the cost, an application is proposed which classifies the regulatory rules that are applicable to that specific industry. Automatic text categorization can be used based on the keywords reorganization. Thus the GUI based can be designed which uses a suitable data mining tool (Mahout) consisting of efficient classification algorithm.

I. LITERATURE SURVEY

In [1] a comparative study of open source tools has been done:

1) WeKa: Weka is a collection of machine learning algorithms for data mining tasks. These algorithms can either be applied directly to a data set or can be called from your own Java code.

2) KEEL: Knowledge Extraction based on Evolutionary Learning is an application package of machine learning software tools designed for providing solution to data mining problems and assessing evolutionary algorithms. It consists collection of libraries for preprocessing and post-processing techniques.

3) R-programming: Revolution is a free software programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

4) KNIME: Konstanz Information Miner, is an open source data analytics, reporting and integration platform. It is based on the Eclipse platform and, through its modular API, and is easily extensible. It provides first- tier support for highly domain-specific data format.

5) RapidMiner: It is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics.

6) ORANGE: Orange is a component-based data mining and machine learning software suite, featuring a visual programming front- end for explorative data analysis and visualization, and Python bindings and libraries for scripting.

In [2], process of classification of text data has been explained:

In [3], types of text mining algorithms (Fig.1):

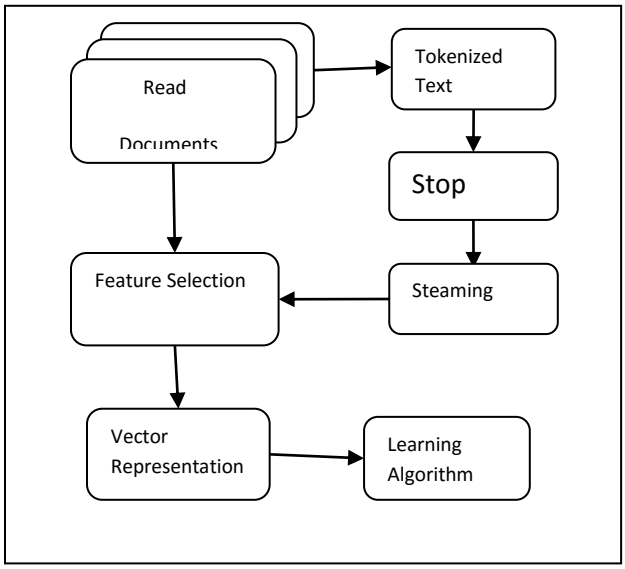

Fig.1 Document Classification Process

- Classification Algorithm:

It is a process of distributing the records into given classes. The attributes of training dataset are used to classify the test data into given classes.

Types of classification models:

1. Decision tree
2. Neural network
3. Generic algorithm

- Clustering Algorithm:

It is a process of dividing the given data into groups of similar objects. Each group is called cluster. Objects in a cluster are similar to each other and dissimilar to objects of other clusters.

Types of clustering models:

1. Hierarchical Methods
2. Partitioning Methods

In [4], A review of text mining is given. It is important to pre-process the text before clustering as mining from pre-processed text becomes easier as compared to natural language documents. Special methods such as filtering and stemming are applied to reduce the dimensionally of the documents words. Irrelevant words from set of all words are removed by filtering. The standard filtering method is stop word filtering or stop words removal. Words like conjunctions, prepositions, articles, etc. are removed by filtering. To find the root/stem of a word stemming technique is used. Stems are considered to be useful for improving retrieval performance because they reduce variants of the same root word to a common concept. Stemming has another effect of reducing the size of the indexing structure because the number of different index terms is reduced.

In [5], Algorithms are explained:

1. The k-means algorithm:

The algorithm operates on a set of d-dimensional vectors, $D = \{xi \mid i = 1, \ldots N\}$, where $xi \in d$ denotes the ithdata point

Step 1: Data Assignment. Each data point is assigned to its closest centroid which results in partioning. Step 2: Relocation of "means". Each cluster represent the center of all data points assigned to it.

2. kNN: k-nearest neighbor classification:

It finds a group of k objects in the training set that are closest to the test object There are three key elements of this approach: a set of labeled objects, a specific metric to calculate distance between objects, and the number of nearest neighbors. To categorize an unlabeled object, the distance of this object to the labeled objects is calculated, its k-nearest neighbors are identified.

3. Naive Bayes:

If we define P(i|x) to be the probability that an object with measurement vector x = (x1,...,xp) belongs to class i, then any function of P(i|x) would make a suitable score. Elementary probability tells us that we can decompose P(i|x) as proportional to f (x|i)P(i), where f (x|i) is the conditional distribution of x for class i objects, and P(i) is the probability that an object will belong to class i. This means that the ratio becomes:

P(1|x) P(0|x) = f (x|1)P(1) f (x|0)P(0) .

[6] is an e-book which explains why Mahout is the best tool for data mining.

Mahout is an open source machine learning library from Apache. As the application gets deployed, there might arise a need of using larger data sets in future. Apache Mahout is used being currently used by a range of many projects working on data mining. It has got an ability to deal with larger data sets. Until and unless the data with double the input and lower capacity is accepted, the system works fine. But if the input is more than the five times of data then a solution must be found for that. This solution is provided by Mahout. It requires resources that takes the computing time and memory less then input data.

[7] Describes about feature selection:

Feature selection is a method of picking only selected attributes which are subset of the training set. It basically makes efficient use of classifier and increases accuracy by reducing noise. It enables faster machine learning as the model becomes simpler and generalized

[8] Describes about preparing data inputs.

1. Collecting the Data:

Incorporating data from many sources usually gives rise to many challenges.

Standard Format:-

There is a standard way of representing datasets. Example: ARFF format.

2. Sparse Data:-

Many attributes might have a value of 0 for most of the instances.

3. Attribute Types:-

The standard format has two basic data types: nominal and numeric. String attributes are mostly nominal and date attributes are numeric, although strings are to be converted into a numeric form such as a word vector.

4. Missing Values:-

Most datasets contain missing values which are normally specified by out-of-range entries like a negative number (e.g., –1) in a numeric field which is normally only positive, or a 0 in a numeric field that can never be 0. For nominal attributes, missing values may be indicated by blanks or dashes.

5. Inaccurate Values:-

It is essential to check the input files carefully for bad attribute values. The data used for mining has almost certainly not been gathered specifically for that purpose. When the data is originally gathered, many of the fields perhaps didn't matter and therefore might have been left blank or unchecked such that it does not affect the original purpose of mining

6. Getting to Know Your Data:-

Get to know your data by the means of graphical visualizations of data such as histograms of nominal attributes and graphs of numeric attributes which makes it easy to identify the outliers and understand the data well. It might represent errors in a data file and identify hidden conventions which will help in coding unusual states. [9] and [10] gives a brief about various clustering algorithms (Table 1):

Cluster Algorithm	Complexity	Capability of Tackling high dimension-al data
K-means	O(NKd)(time) O(N+K)(space)	No
Fuzzy c-means	Near O(N)	NO
Hierarchical Clustering	$O(N^2)$(time) $O(N^2)$(space)	No
CLARA	$O(K(40+K)^2+K(N-K))^+$ (time)	No
BIRCH	O(N)(time)	No

Table 1: Time complexities of algorithms

1. Simple K-Means Clustering

In K-Means algorithm, clusters are formed from items by taking mean of some attribute of items.

Algorithm:

- Assign initial value to means.
- Add the next element to the cluster with closest mean.
- Calculate the mean again after each element is added.

This algorithm is not suitable for large data sets since it takes more time to form clusters.

2. Farthest First Clustering

This algorithm places the center of next cluster at the next point from current cluster, where this point lies within the data area. The farther points are clustered first then the nearest points are clustered. This speeds up the clustering process in situation like reassignment.

3. Make Density Based Clustering

This algorithm is used in cases where clusters are irregular or outliers are encountered or in noisy situation. The points lying in same area and with same density are clustered together.

II. PROPOSED FRAMEWORK

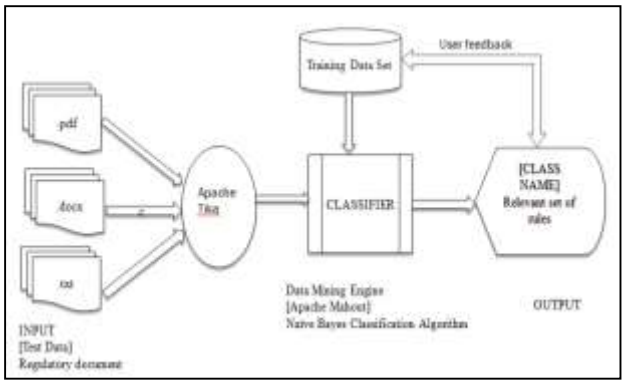

Fig 2: System Architecture

An Input file consist of regulatory rules. It can be of any format (.pdf, .txt, .docs etc). This input file is passed through Apache Tika in which, only the text is extracted and forwarded to the classifier. The classifier takes this extracted text as input from Tika. It then classifies the text based on a predefined classifier model constructed with the help of training data set and displays the rules (Fig.2).

III. PROPOSED ALGORITHM AND TOOL

Algorithm (Fig.3):

1. Accept an input file (Fig. 4).
2. Extract text using Apache Tika.
3. Classify the file using training data set.
4. Display the rules and their belonging classes (Fig.5).
5. The selected class details would be given as feedback to database.
6. Display final result.

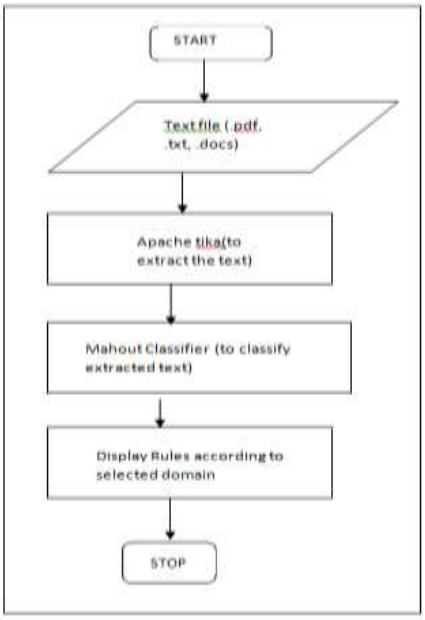

Fig 3: Flow-Chart of the system.

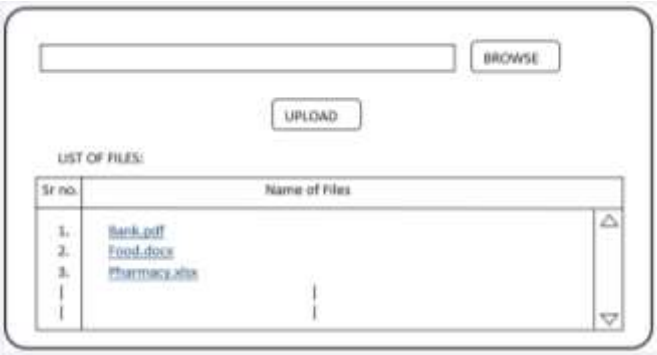

Fig 4: Proposed GUI I

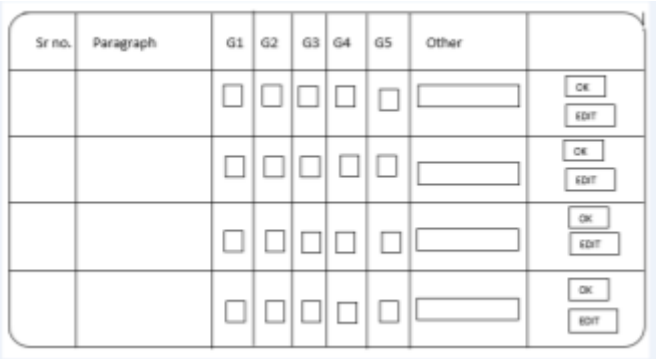

Fig 5: Proposed GUI II

Tool:

Name: Apache Mahout

Description: It produces free implementations of distributed and otherwise scalable machine learning algorithms on the Hadoop platform

Advantages: Faster interaction with database.

Efficient use of algorithms

IV. EXPECTED RESULTS

1) Input: File based on regulatory rules.

2) Intermediate Results

a) MD5 result

b) Rules classified into classes.

3) Output: Final classification after user feedback.

V. ADVANTAGES

1. Helps to store and retrieve information easily- The categories created become storage units, to name or label these rules that come under each labeled category, so that it is flexible to retrieve them by searching their location.

2. Classification reduces complexity- It provides a way for grouping to understand the complexity better.

3. Efficient use of algorithms-The use of data mining tools allows efficient use of algorithms. The algorithm chosen requires only one pass over data.

4. The algorithm only requires a small amount of training data to estimate the parameters necessary for classification.

5. Algorithm is robust enough to ignore the deficiencies.

VI. APPLICATIONS

1. Industries can use this application for searching rules related to their domain.

2. Document Processing and Analysis.

VII. CHALLENGES

The feature vector or other structure which is used to indicate a document has complex semantics of natural language. The features must cover a wide range of class definitions. Some classifiers in apache mahout won't be able to go beyond 85% accuracy no matter how much training they get.

VIII. CONCLUSION AND FUTURE WORK

Naïve Bayes is the best suited algorithm for text classification for medium data sets to large data sets. A Study of formulating the input for training data set is completed. Hadoop can be later on integrated in the same application.

IX. REFERENCES

1. Kalpana Rangra, Dr. K. L. Bansal," Comparative Study of Data Mining Tools", IJARCSSE, Volume 4, Issue 6, June 2014 ISSN: 2277 128X

2. Aurangzeb Khan, BaharumBaharudin, Lam Hong Lee*, Khairullahkhan, "A Review of Machine Learning Algorithms for Text-Documents Classification", JOURNAL OF ADVANCES IN INFORMATION TECHNOLOGY, VOL. 1, NO. 1, FEBRUARY 201

3. Bhumika, Prof Sukhjit Singh Sehra, Prof AnandNayyar, "A REVIEW PAPER ON ALGORITHMS USED FOR TEXT CLASSIFICATION", International Journal of Application or Innovation in Engineering & Management (IJAIEM) Web Site: www.ijaiem.org

Email:editor@ijaiem.org,editorijaiem@gmail.com, Volume 2, Issue 3, March 2013 ISSN 2319 – 4847

4. Greeshma RG, Smitha ES "A review on mining text data with auxiliary attributes", July 2015, Volume 2, Issue 6 JETIR (ISSN-2349-5162)

5. Xindong Wu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg, "Top 10 algorithms in data mining", KnowlInfSyst (2008) 14:1–37 DOI 10.1007/s10115-007-0114-2 SURVEY PAPER, Received: 9 July 2007 / Revised: 28 September 2007 / Accepted: 8 October 2007 Published online: 4 December 2007 © Springer-Verlag London Limited 2007

6. Sean Owen, Robin Anil, Ted Dunning, Ellen Friedman, "Mahout in Action", ©2012 by Manning Publications Co., www.manning.com.

7. Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze "Introduction to Information Retrieval" ", ©2013.

8. Ian H. Witten ,Eibe Frank , Mark A. Hall ," DATA MINING practical Machine Learning Tools and Techniques (3rd edition) "

9. Rui Xu and Donald Wunsch II, "Survey of Clustering Algorithms", IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005 645.

10. S. Revathi, Dr.T.Nalini, "Performance Comparison of Various Clustering Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Available online at: www.ijarcsse.com , Volume 3, Issue 2 February 2013 ISSN: 2277 128X