MANTECH Publications

Journal of Computer, Internet and Network Security
Volume 5 Issue 2
ISSN: 2457-0176 (Online)

"Descriptive Clustering of Documents on the Basis of Predictive Network"

Prof. Suraj Shivaji Bhoite, Prof. Swapnali K. Londhe

Assistant Professor

Department of Computer Science Engineering

S.M.S.M.P.I.T.R, Akluj, Solapur

Corresponding Author's Email id: *surajbhoite19@gmail.com, tejswapn@gmail.com*

Abstract

The mass of data obtainable, merely finding the IR is not only assignments of automatic text classification methods. Instead, the automatic text classification methods are supposed to retrieve the RI as well as organize according to it is a degree of relevancy with the given query. The main problem in organizing is to classify which document is relevant & which are relevant. The Automated text classification consists of automatically organizing clustered data. We propose an automatic method of text classification using Machine Learning based on the disambiguation of the meaning. We use the word net word embedding algorithm to eliminate the ambiguity of words so that each word is replaced by its meaning in context. The closest ancestors of the senses of all the undamaged words in a given document are selected as classes for the specified document.

Keywords*: - Machine Learning, Logistic Regression, Descriptive Clustering.*

INTRODUCTION

The mass of data obtained by us increases. This data would be irrelevant if our capacity to productively get to didn't increment too. For most extreme advantage, there is a need for devices that permit look, sort, list, store and investigate the accessible information. One of the hopeful regions is the automatic text classification. Envision ourselves within sight of an impressive number of texts, which are all the more effectively available on the off chance that they are composed into classes as per their topic.

Obviously, one could request that humans read the text and arrange them physically. This assignment is hard if done on hundreds, even a huge number of texts. Thus, it appears to be important to have a computerized application, so automatic text categorization is presented here. Growing the number of Data Mining presentations includes the complex & structured type of Data & Requires the use of sensitive language.

Unfortunately Existing approach, especially by using Logic Programming methods, often suffer is not low scalability. When allocating with difficult DataBase schemas but also insufficient predictive performance usage noisy value in real-life applications. Though leveling approaches are wont to need significant time & effort for the data conversion, the outcome in behind the compressed representations of the standardized Database & produce is extremely large with a huge number of extra attribute value & frequent NULL values. In this system, the result is problems that have prevented a diverse application of multi Relational Mining and the Data Mining community challenge. This article introduces a Descriptive clustering methodology where flattening is required to bridge the gap between propositional learning algorithms and relational.

In the proposed methodology, Data Analysis Technique, clustering can classify a subset of information examples with mutual characteristics. The operator can explore the information by examining the user can search the data by using a selected example in the group instead of relatively examining the instances of the complete data set. This allows users to focus efficiently on large relevant subsets Data sets, in particular for document collections. In particular, the descriptive grouping consists of an automatic combination set of parallel instances in the cluster and repeatedly generates for the explanation or synthesis that can be interpreted by man for each group. The description of each cluster allows a user to determine the grouping's relevance to observe its content in the text documents. Quality is the grouping is important, so that it is aligned with the idea of the likeness of the user, but it is similarly imperative to deliver a user with a brief & useful instant that correctly replicates the constant of the cluster.

A. Motivation

In the introductory part for the study of Text Classification, their application, which algorithm used for that and the

different types of model, I decided to work on the Text Classification, which is used for data analysis lot of work done on that application and that the technique used for that application is Text Classification using traditional data mining algorithms and not worked on word sense technique. Approaches to state of the art to classify data it can be used to find the subset of data examples. However, they suffer from low accuracy.

RELATED WORK

A literature survey is the utmost significant steps in any kind of investigation. Already start developing, we need to study the previous papers of our domain, which we are working on and on the basis of the study, we can predict or generate the drawback and start working with the reference of previous papers. In this section, we briefly analyse the related work on Text classification & their different techniques.

J – T. Chien, describe the "Hierarchical Theme & Topic Modeling," in that Taking into account hierarchical data sets in the body of text, such as words, phrases, and documents, they perform Structural Learning and they deduce hidden theme, and themes for sentences and words from a group of documents, correspondingly. The association of the arguments &arguments in dissimilar Data groups are discovered through an unverified procedure without preventive the number of the clusters. A tree branching process is presented to draw the scopes of the topic for different phrases. They build a Hierarchical theme and a thematic model, which flexibly represents heterogeneous documents using non-parametric Bayesian parameters. The thematic phrases and the thematic words are extracted. The proposed method is evaluated as active for the construction of a Semantic Tree Structure of the corresponding Sentences and Words. The superiority of the use of the tree model for the variety of sensitive phrases for the summary of documents is illustrated [1].

M. D Amico, & Bemardini, describes the "Full-Subtopic Retrieval with key-Phrase Based Search Results clustering," study of this problem of restoring many documents that are relevant to individual sub-topics give the Web query, called "full child retrieval." Solution to this problem, they are using a new algorithm for grouping search results that generate clusters labelled with key phrases. The key phrases are removed general suffix tree created by the results and merge through a hierarchical agglomeration procedure improved grouping. They also introduce a

new measure to evaluate the performance of full recovery sub-themes, namely "look for secondary arguments length under the sufficiency of k documents." They have used a test collection specifically designed to evaluate the recovery of the sub-themes. They have found that our algorithm has passed both other clustering algorithms of present research outcomes as a method of redirecting search results underline the diversity of results. [2].

T. Kohonen, S. Kaski, J. Honkela, A. Saarela, K. Lagus, "Self Organization of a Massive Document," in this paper defines the execution of the system that can organize large collections of documents based on written parallels. It's based upon the self-organized map (SOM) algorithm. Like the feature courses for documents, the arithmetical symbols of their words are used. The important objective of this work was to resize the SOM Algorithm in order to handle large amounts of high-dimensional data. In practical experimentation, they mapped 6 840 568 patent abstracts in a SOM of 1.002.240 nodes. As characteristic vectors, we use 500 stochastic data vectors as unsystematic estimates of histograms of biased words [3].

S. Roy, K. Single, R. Krishnpuram, K. Kummamuru describe the "A Hierarchical Monothetic Document clustering Algorithm For Summarization & Browsing Search Result," this paper Organizing Web search results in a hierarchy of themes and secondary theme make it easy to explore the collection and position the results of awareness. They are proposing a new hierarchical monarchic grouping algorithm to construct a hierarchy of topics for a collection of search results retrieved in response to a query. At all levels of the hierarchy, the new algorithm increasingly finds problems in order to maximize coverage and maintain the distinctiveness of the topics. They discuss the algorithm proposed as Discover. The evaluation of the quality of a hierarchy of subjects is not a trivial task; the last test is the user's judgment. An algorithm is a bit more computationally than other algorithms. It generates better hierarchies. The study also shows that the proposed algorithm is superior to other algorithms as a tool for summary and navigation [4].

D. Wunsch & R. Xu describes the "Survey of Clustering Algorithms," in that Data Analysis shows of crucial role inconsiderate the several occurrences. Conglomerate Analysis, a primitive

examination with little or no earlier knowledge, contains the research developed in a Wide Variety of communities. Diversity, on the one hand, provides us with many tools. On the other hand, the profusion of options causes confusion. They have examined the grouping Algorithm for the Data-Sets that appear in indicators, Computer Science & Machine Learning, and they illuminate their uses in some reference Data-Set, the problematic of street vendors & Bioinformatics, and one area that attracts intense exertions. Various closely associated areas, immediacy measurement and Cluster authentication. [5].

D. Heckeman, J. Platt, M. Sahami & S. Dumais, "Inductive Learning Algorithm & Representation for Text Categorizations," in Text categorization is the task of Natural Language Texts to one or more predefined forms based on their content is a significant component in many data association and managing tasks. They are associate to the efficiency of five dissimilar Automatic Learning Algorithm for text categorization in terms of Learning rapidity, present organization speed and organization accurateness. They are also examining training set size and alternative document representations. Very accurate text classifiers can be learned automatically from training examples. [6].

R. Kohavi and G. John, describe the "Wrappers for Feature Sub-set Selection," In this paper, feature sub-set selection problematic, a Learning Algorithm is challenged with the problem of selecting a Relevant Sub-set of a Feature upon which to focus it is attention. A feature subset selection technique should consider how the algorithm and the training set interact to attain the best possible performance with a specific learning algorithm on a specific training set. They discover the relative optimal feature, sub-set selection and significance. Their wrapping method is searching for an optimal feature subset personalized to a specific algorithm and domain. An important development in accuracy is achieved for some datasets for the two relations of induction algorithms used: decision trees and Naive-Bayes [7].

C. Zhai & X. Shen, describes the "Automatic Labeling of Multinational Topic Models," In this paper, they suggest probabilistic methods to repeatedly labeling multinomial topics copies are objective way. They are cast of labeling problem is an optimization problem involving reducing word distributions and

exploiting mutual info between a label and a topic model. Experiments with user studies have been done on two text data sets with different genres. These system results are the proposed labeling methods that are quite effective in generating meaningful and useful labels for interpreting the discovered topic models. Their methods are universal and can be applied to labeling topics learned through the kinds of topic models such as PLSA, LDA, and their variations [8].

OPEN ISSUE

The work has been done in this planning because it is general usage & requests. In this Part, some of the methods that have been executed to realize a similar purpose are declared. It's about works that are majorly separated by the algorithm for Text Classification. In another research, accessing the relevant data from a mass of data is a very difficult and time-consuming task as the mass of Data increases because of the digital world. The mass of data obtainable to us increases. This data would be irrelevant if this information would be irrelevant if our capacity to proficiently access did not increase as well. Automated text classification provides us with maximum benefit that allows us to search, sort, index, store, & analyze the available data. It also allows us to find in desired information in a sensible period.

PROPOSED APPROACHES

We propose automatic text classification using TFIDF, Word sense word embedding algorithm and K-means clustering classification machine learning algorithm, firstly apply the pre-processing technique and remove unwanted data, then calculate TFIDF and Semantic relations of each and every word and classify data using Machine Learning Algorithm on conference papers dataset.

Steps:

1. Training Module

a) Load Dataset
b) Extract Dataset
c) Remove Stop Word
d) Apply Stemming
e) Calculate TF-IDF Score
f) Train Data

2. Testing Module

a) Load Testing PDF Papers
b) Extract Data
c) Remove Stop Word
d) Apply Stemming
e) Calculate TF-IDF Score
f) Extract Feature
g) Classify Data using Linear Regression
h) View Analysis Graph

System Architecture

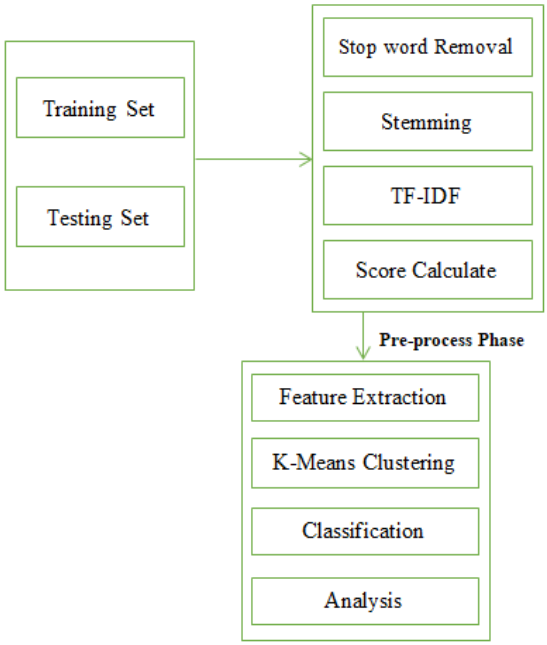

Fig1. System Architecture

Algorithms

1. **Reprocessing Algorithms**

a) ***Stop Word Removal-*** This technique removes stop words like **is, are, the, an, in, they, but etc.**

b) ***Tokenization-*** This technique removes Special characters and images.
For e.g. - #, *, $, @ etc.

c) ***Stemming Remove-*** This algorithm process is semantic standardization, in which the different forms of the word the reduced to the mutual form. This technique removes suffix and prefix and Find Original word.
For e.g. - i. Played - play &
ii. Clustering – cluster

2. **TFIDF Algorithm**

The *tf - idf* score are term *Iijis* calculated by the Term Frequency & IDF. The TF & IDF is the fundamentals of the greatest general term weighting system in IR. The *tf – idf* total of *Iij*, *tf – idf(Iij)*, this formula are calculated as the *tf – idf (Iij)* –

Tf – idf (Iij) log (tf (Iij ,dj) + 1) * log (| D | / 1 +df (Iij , D))

MATHEMATICAL MODEL

Let us consider S is a System for Document Classification.

S *= {I, F, O}*

Where,

I = Text Dataset Uploaded by the User

F= Functions Implemented to get the Output

O = Output i.e. Document Clustering

Input:-

Classify the Inputs:

Set,

F*= {F1, F2, F3……..Fn}*

Where *F* is a Set of the Function to Execute Commands

I*= {I1, I2, I3……In}*

Where *I* is a Set of Input to the Function Set

O *= {O1, O2, O3……..On}*

Where *O* is Set of Output from the Function Set

RESULT

In a tentative outcome, we evaluate this system which is based on the student conference papers. Dataset is available on the internet. We compare the accuracy of the existing system results with the proposed system.

The result evaluation is given below:-

TP: True Positive (Properly Expected No. of Instance)

FP: False Positive (wrongly Expected No. of Instance)

TN: True Negative (Properly Expected No. of Instance as not required)

FN: False Negative (Wrongly Expected No. of Instance as not required)

We can Calculate Four Measurements:-

Accuracy: TP+TN/TP+FP+TN+FN

Precision: TP/TN+FP

Recall: TP/TP+FN

F1-Measure: 2*Precision *Recall / Precision +Recall.

A. Comparison Graph

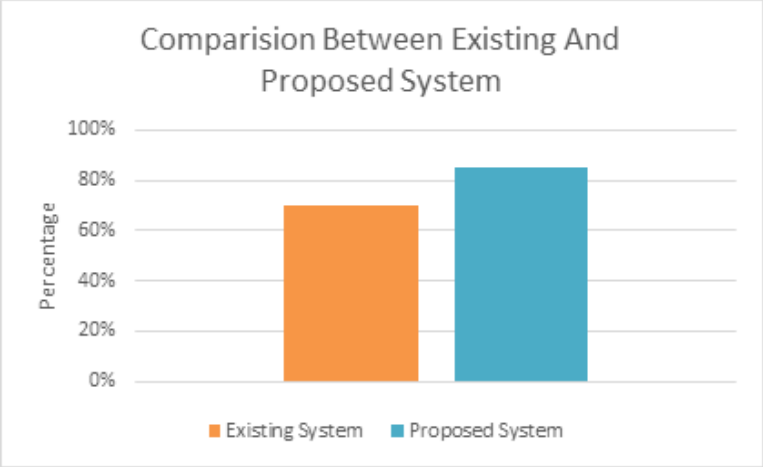

Fig. 2 Comparison Graph

B. Comparison of Table

Table 1: Comparative Result

Sr. No	Existing Method Result	Proposed Method Result
1	65%	84%

C. Accuracy Graph

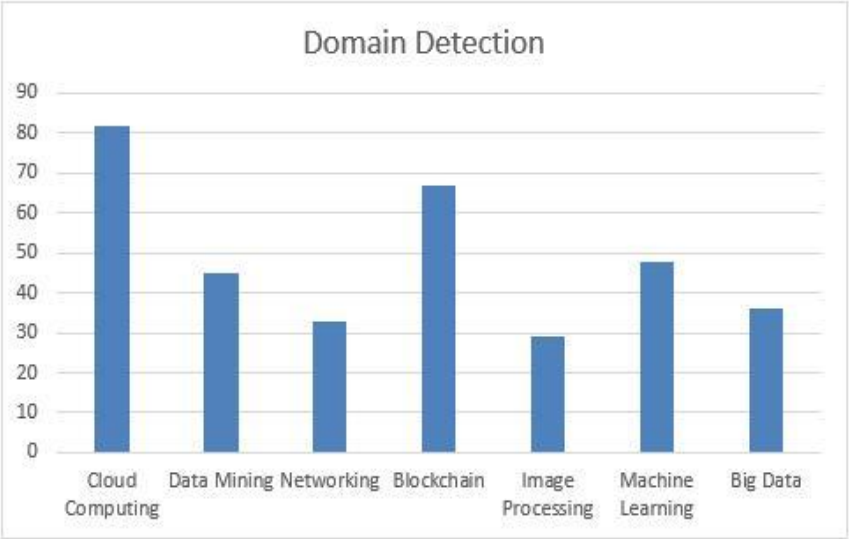

Fig. 3 Accuracy Graph

D. Domain Detection Result

Table 2: Domain Detection Result

Sr. No.	Domain Name	Paper Count
1	Cloud Computing	82
2	Data Mining	45
3	Networking	34
4	Blockchain	65
5	Image Processing	29
6	Machine Learning	46
7	Big Data	36

CONCLUSION

Proposed descriptive Clustering as two coupled predictions activity chooses a grouping are predictive of type and Predictions of Cluster task of the sub-set of features. Use of predictive performance a goal standard, Descriptive Clustering constraints the No. of Clusters & the No. of Functions each Clusters: they are selected from the model selection. With the outcome, each group is a minimum subset of a feature required to expect, if instance, belongs to the cluster. Our suggestion is that even an operator will be clever to forecast membership in this group of papers using the descriptive features selected by the algorithm. Given some relevant requirements, the operator can speedily identify clusters that probably contain applicable papers.

REFERENCES

I. J – T Chien, "Hirarchical Theme & Topic Modeling," IEEE Trans. 2016.

II. Bemardini, Carpineto, & M.D Amico, "Full- Subtopic with Key phrase – Based Searching Results Clustering," IEEE 2009.

III. T. Kohonen, S. Kasski, J. Salojarvi, V. Paatero, K. Lagus, & A Saarela, "Self-Organization of a Massive Document Collection," IEEE Trans. 2000.

IV. R. Lotlikar, K. Kummamuru, K. Singal, R. Krishnapuram, "A Hirarchical Monothetic Document Clustering Algorithm for Summarization & Browsing Search Result, "Proc. Int. Conf 2004.

V. Wunsch & R. xu, "Survey of Clustering Algorithms," IEEE Trans. 2005.

VI. D. Heckeman, J. Platt, S. Dumains & M. Sahami, "Inductive Learning Algorithm & Representations for Text Categorization," in Proc. Int. Conf.1998.

VII. H John, & R. Kohavi, "Wrappers for Feature Subset Selection," Artif. Intell. 1997.

VIII. Zhai & Q. Mei, "Automatic Labeling of Multinational Topic Models," Pro. AcmSigkddInt .Conf 2007.

AUTHORS PROFILE

Prof. Suraj S. Bhoite

Working as an Assistant Professor in SMSMP Institute of Research and Technology, Akluj, Solapur, Maharashtra.

Prof. Swapnali K. Londhe

Working as an Assistant Professor in SMSMP Institute of Research and Technology, Akluj, Solapur, Maharashtra.