ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

# Predictive Descriptive Clustering using Semantic Score

Miss.Panchashila Mokal , Prof. Anilkumar Brahmane

Department of Computer Engineering, SRES COE, Kopargaon, India

Department of Computer Engineering SRES COE, Kopargaon, India

**ABSTRACT:** The descriptive document clustering consists of automatically text classification and generating a descriptive summary for each group. The description should inform a user about the contents of each cluster without further examining the specific instances, allowing the user to quickly search for the relevant clusters. Consistently the mass of data accessible, simply finding the significant data isn't the main undertaking of programmed content characterization frameworks. Rather the automatic text classification systems are supposed toretrieve the relevant information as well as organize according to its degree of relevancy with the given query. The main problem in organizing is to classify which documents are relevant and which are irrelevant. The Automated content grouping consists of automatically organizing clustered data. We propose an programmed strategy for content characterization using machine learning based on the disambiguation of the meaning of the word we utilize the word net word installing calculation to dispose of the equivocalness of words with the goal that each word is supplanted by its importance in setting. The nearest precursors of the faculties of all the intact words in a given record are chosen asclasses for the pre defined report.

**KEYWORDS:** Document clustering, feature selection, model selection, machine learning

## I. INTRODUCTION

This information would be irrelevant if our ability to productively get to did not increment too. For most extreme advantage, there is need of devices that permit look, sort, list, store and investigate the accessible information. One of the promising region is the automatic text categorization. Envision ourselves within the sight of 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 human read the text and arrange them physically. This assignments is hard if done on hundreds, even a huge number of texts. Thus, it appears to be important to have a computerized application, so here automatic text categorization is presented. An increasing number of data mining applications involve the analysis of complex and structured types of data and require the use of expressive pattern languages. Many of these applications cannot be solved using traditional data mining algorithms. This observation forms the main motivation for the Linear Regression.

Unfortunately, existing upgrading approaches, especially those using Logic Programming techniques, often suffer not only from poor scalability when dealing withcomplex database schemas but also from unsatisfactory predictive performance while handling noisy or numeric values in real-world applications. However, flattening strategies tend to require considerable time and effort for the data transformation, result in losing the compact representations of the normalized databases, and produce an extremely large table with huge number of additional attributes and numerous NULL values (missing values). As a result, these difficulties have prevented a wider application of multi relational mining, and post an urgent challenge to the data mining community. To address the above mentioned problems, this article introduces a Descriptive clustering approach where neither upgrading nor flattening is required to bridge the gap between propositional learning algorithms andrelational.

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

In Proposed approach, Data analysis techniques, such as clustering it can be used to identify subsets of data instances with common characteristics. Users can explore the data by examining some instances in each group instead ofrather than examining the instances of the complete data set. This allows users to focus efficiently on large relevant subsetsData sets, in particular for document collections. In particular, the descriptive grouping consists of automatic grouping sets of similar instances in clusters and automatically generate a description or a synthesis that can be interpreted by man for each group. The description of each cluster allows a user determine the relevance of the group without having to examine its content For text documents, a description suitable for each group can be a multi-word tag, an extracted title or a list of characteristic words . The quality of the grouping it is important, so that it is aligned with the idea of likeness of the user, but it is equally important to provide a user with a brief and informative summary that accurately reflects the contents of the cluster.

## II. LITERATURE SURVEY

J.-T Chien,describe the Hierarchical theme and topic modeling ,In that Taking into record progressive informational indexes in the assortment of content, for example, words, expressions and archives, we perform auxiliary learning and we find inert subjects and topics for sentences and words from an accumulation of reports, separately. The connection among contentions and contentions in various information groupings is investigated through an unsupervised strategy without constraining the quantity of bunches. A tree stretching process is introduced to draw the extents of the subject for various expressions. They fabricate a various leveled subject and a topical model, which adapt ably speaks to heterogeneous records utilizing non-parametric Bayesian parameters. The topical expressions and the topical words are removed. In the investigations, the proposed technique is assessed as successful for the development of a semantic tree structure for the relating sentences and words. The prevalence of the utilization of the tree demonstrate for the determination of expressive expressions for the outline of reports is represented[1].

Bernardini, C. Carpineto, and M.DAmico, describe the Full-subtopic retrial with keyphrase based search result clustering,in that they Consider the issue of reestablishing numerous reports that are applicable to the individual sub-points of a given Web inquiry, called ”full youngster recovery”. To take care of this issue, they present another calculation for gathering indexed lists that produces groups marked with key expressions. The key expressions are removed summed up postfix tree made by the query items and converge through a progressive agglomeration system enhanced gathering. They additionally acquaint another measure with assess the execution of full recuperation sub-subjects, in particular ”search for optional contentions length under the adequacy of k records”. they have utilized a test accumulation explicitly intended to assess the recuperation of the sub-subjects, they have discovered that our calculation has passed both other grouping calculations of existing exploration results as a technique for diverting list items underline the decent variety of results (in any event for k¿ 1, that is the point at which they are keen on recouping more than one pertinentre port by sub-topic)[2].

R. Krishnapuram, describe the A hierarchical monothetic document clustering algorithm for summarization and browsing search result,in that Organizing Web search results in a hierarchy of topics and secondary topics makes it easy to explore the collection and position the results of interest. In this paper, they propose a new hierarchical monarchic grouping algorithm to construct a hierarchy of topics for a collection of search results retrieved in response to aquery. At all levels of the hierarchy, the new algorithm progressively identifies problems in order to maximize coverage and maintain the distinctiveness of the topics. They refer to 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. They have used various objective measures, such as coverage and application time for an empirical comparison of the proposed algorithm with two other monoteticgrouping algorithms to demonstrate its superiority. Although our algorithm is a bit more computationally than one of the algorithms, it generates better hierarchies. Our user studies also show that the proposed algorithm is superior to other algorithms as a toolforsummaryandnavigation[4].

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

R. Xu and D. Wunsch, describe the Survey of clustering algorithms, in that Data analysis plays an indispensablerole in understanding the various phenomena. Conglomerate analysis, primitive exploration with little or no previous knowledge, consists of 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 algorithms for the data sets that appear in statistics, computer science and machine learning and they illustrate their applications insome reference datasets, the problem of street vendors and bioinformatics, and a new field that attracts intense efforts. Various closely related topics, proximity measurement and cluster validation are also discussed[5].

S. Dumais, J. Platt, D. Heckerman, and M. Sahami,describe the Inductive learning algorithms and representations for text T Kohonen, S.Kaski, K.Lagus, J.Salojarvi, J.Honkela, V.Paatero, categorization, in that Text categorization theassignment of and A.Saarela describe the Self-Organization of massive document collection,in this they depicts the usage of a framework that can sort out vast accumulations of reports dependent on printed likenesses. It depends on oneself sorted out guide (SOM) calculation. Like the element vectorsfor records, the measurable portrayals of their vocabularies are utilized. The fundamental goal of our work was to resize the SOM calculation so as to deal with a lot of high-dimensional information. In a reasonable analysis, they mapped 6 840 568 patent modified works in a SOM of 1.002.240 hubs. As trademark vectors, we use vectors of 500 stochastic figures acquired as arbitrary projections of histograms of weighted words[3].

K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and natural language texts to one or more predefined categories based on their content is an important component in many information organization and management tasks. They compare the effectiveness of five different automatic learning algorithms for text categorization in terms of learning speed, real-time classification speed and classification accuracy. They also examine training set size, and alternative document representations. Very accurate text classifiers can be learned automatically from training examples. Linear SupportVector Machines (SVM) are particularly promising because they are very accurate, quick to train and quick to evaluate[6].

R. Kohavi and G. H. John, describe the Wrappers for feature subset selection, In that the feature subset selection problem, a learning algorithm is faced with the problem of selectinga relevant subset of features upon which to focus its attention, while ignoring the rest. To achieve the best possible performance with a particular learning algorithm on a particular training set, a feature subset selection method should consider how the algorithm and the training set interact. They explore the relation between optimalfeature subset selection and relevance. Our wrapper method searches for an optimal feature subset tailored to a particular algorithm and a domain. They study the strengths and weaknesses of the wrapper approach and show a series of improved designs. They compare the wrapper approach to induction without feature subset selection and to Relief, a filter approach to feature subset selection. Significant improvement in accuracy is achieved for some datasets for the two families of induction algorithms used: decision trees and Naive-Bayes[7].

T.Kohonen,S.Kaski,K.Lagus,J.Salojarvi,J.Honkela, Paatero, and A. Saarela, describe the Self-organization of a massive document collection, This paper describes the implementation of a system that is able to organize vast document collections according to textual similarities. It is based on the self-organizing map (SOM) algorithm. As the feature vectors for the documents statistical representations of their vocabularies are used. The main goal in our work has been to scale up the SOM algorithm to be able to deal with large amounts of high-dimensional data. In a practical experiment we mapped 6 840 568 patent abstracts onto a 1 002 240-node SOM. As the feature vectors we used 500-dimensional vectors of stochastic figures obtained asrandom projections of weighted word histograms[8].

Q. Mei, X. Shen, and C. Zhai, describe the Automatic labeling of multinomial topic models, In this paper, they propose probabilistic approaches to automatically labelling multinomial topic models in an objective way. They castthis labelling problem as an optimization problem involving minimizing Kullback-Leibler divergence between word distributions and maximizing mutual information between a label and a topic model. Experiments with user study have

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

been done on two text data sets with different genres. The results show that the proposed labelling methods arequite effective to generate labels that are meaningful and useful for interpreting the discovered topic models. Our methodsare general and can be applied to labelling topics learned through all kinds of topic models such as PLSA, LDA, and their variations[9].

K. Lagus and S. Kaski, describe the Keyword selection method for characterizing text document maps, in that Char- acterization of subsets of data is a recurring problem in data mining. They propose a keyword selection method that can be used for obtaining characterizations of clusters of data whenever textual descriptions can be associated, with the data. Several methods that cluster data sets or form projections of data provide an order or distance measure of the clusters. If such an ordering of the clusters exists or can be deduced, the method utilizes the order to improve the characterizations. The proposed method may be applied, for example, to char- acterizing graphical displays of collections of data ordered

e.g. with the SOM algorithm. The method is validated using a collection of 10,000 scientific abstracts from the INSPEC database organized on a WEBSOM document map[10]

## III. PROPOSED METHODOLOGY

In Proposed System training is creation of train data set using which classification of unknown data in predefined categories is done. Here a learning system is created using regression. It is a supervised learning where unlabeled data is classified using labelled data. Training data is always alabelleddatasetbasedonitsfeatures.

Project had considered no of scientific papers form dif- ferent publication of different domains for creating training dataset. These papers are input for creating training dataset. This input is first preprocessed and most informative features are extracted using TF/IDF algorithm and word embedding sematic score algorithm. Ten different domains from market are identified and then extracted feature and have to put to corresponding domain where each domain is considered as one class that which is used for labeling test dataset in testing part and features are considered as nodes. Once training part is completed, all features of respective domains are getupdated in corresponding tables indatabase.

### *A. Architecture*

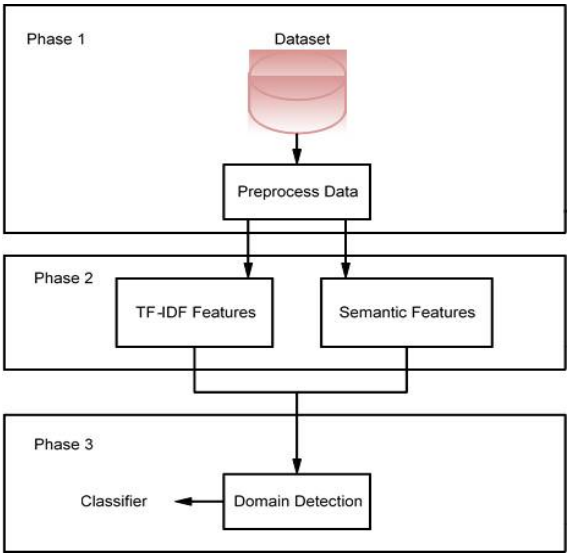


Fig. 1. System Architecture

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

*B. Algorithms*

**Preprocessing Algorithms:**

Stop word Removal-This technique remove stop words like is, are,they,but etc.
Tokenization-This technique remove Special character and images.
Stemming remove suffix and prefix and Find Orignalword for e.g.- 1. playedplay
2.Clustering - cluster

**TFIDF Algorithm**

The tfidf score for term Iij is calculated using the term frequency and the document frequency. Term frequency (tf) and inverse document frequency (idf) are the foundations of the most popular term weighting scheme in IR. The tfidfscore of Iij ,tfidf(Iij).

**Semantic Score Calculation**

In semantic Calculation, the keywords that are selected from the preprocessing techniques are applied to the word net ontology to extract the semantic relation of every keyword. Asynonym is a word, which can be used to substitute another word without a change in the meaning of the words based on the semantic processing, unique keywords are selected in association with the extracted synonym words. The semantic processing is the effective way of text classification with robustness, reliability, and effectiveness. The organizational diagram of these mantic keyword processing

*C. Hardware and SoftwareRequirements*

**Hardware Requirements:**

1. Processor - PentiumIII
2. RAM - 2GB(min)
3. Hard Disk - 20GB
4. Key Board - Standard WindowsKeyboard
5. Mouse - Two or Three ButtonMouse
6. Monitor -SVGA

**Software Requirements:**

1. Operating System -Windows
2. Application Server - ApacheTomcat
3. Coding Language - Java1.8
4. Scripts -JavaScript.
5. Server side Script - Java ServerPages.
6. Database - My SQL5.0
7. IDE -Eclipse

*D. Mathematical Model*

To decide which terms are minor or major information elements in the document, term weighting schemes using three measures are introduced. The two measures are (1) tfidfScore, and (2) Semantic Score First, the tfidf score for term I is calculated using the term frequency and the document frequency. Term frequency (tf) and inverse document frequency (idf) are the foundations of the most popular term weighting scheme in IR. The tfidf score of Iij , tfidf(Iij), is computed as:

Tf-Idf(Iij) = Log (tf (Iij , dj) + 1) * Log (—D—/1 + df (Iij, D))

Where tf (Iij , dj) is the frequency of term Iij within document j and df (Iij , D) is the no. of documents that contain term Iijin the document collection D. Thus, terms with a high tfand lowdfwillgethightf-idfscores.

$$\text{tf}-\text{idf}\left(I_i^j\right) \quad \log\left(tf\left(I_i^j, d_j\right)+1\right) \times \log\left(|D|/1+df\left(I_i^j, D\right)\right)$$

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

**Classification**

Classification using machine learning.

Finally classification using machine learning framework.

1. Given training dataset D which consists of documents belongingtodifferentclasssayClassAandClassB
2. Calculate the prior probability of class A=number of objectsofclassA/totalnumberofobjects

Calculate the prior probability of class B=number of objects ofclassB/totalnumberofobjects

3. Find NI, the total no of frequency of each class Na=thetotalnooffrequencyofclassA

Nb=thetotalnooffrequencyofclassB

4. Find conditional probability of keyword occurrence given aclass:

P (value 1/Class A) =count/ni (A) P (value 1/Class B) =count/ni (B) P (value 2/Class A) =count/ni (A) P (value 2/Class B) =count/ni (B)

..

..

..

P (value n/Class B) =count/ni (B)

5. Avoid zero frequency problems by applying uniform distribution
6. ClassifyDocumentCbasedontheprobabilityp(C/W)

a. Find P (A/W) =P (A)*P (value 1/Class A)* P (value 2/ClassA).

P(value n /Class A)

b. Find P (B/W) =P (B)*P (value 1/Class B)* P(value 2/Class B).

P(value n /Class B)

7. Assigndocumenttoclassthathashigherprobability.

## IV. RESULT ANDDISCUSSION

In experimental results, we evaluate the proposed system on student conference papers datasets this available on internet. We compare the accuracy of existing system results with proposed system and doamin detection of the papers.

| Sr. No. | Existing(Bag of Words) | Proposed(Semantic Relation) |
|---|---|---|
| 1 | 69% | 82% |

Table 1:Comparative Result

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

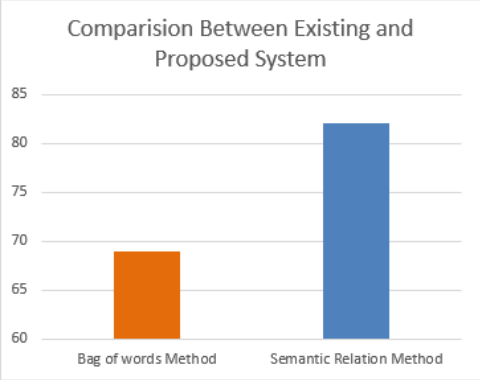


Fig. 2. Accuracy Graph

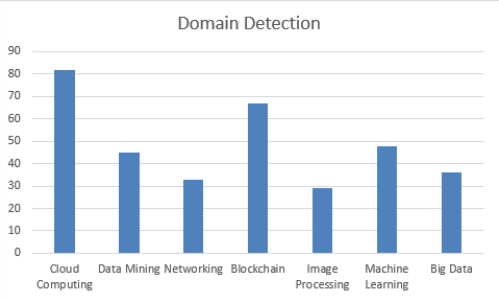


Fig. 3. Accuracy Graph

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

Table 1:Domain Detection Result

## V. CONCLUSION

In this paper, a text classification system was proposed to classify documents in the database based on the subject's label. The proposed classification methodology used the semantic score processing in the extraction of the

ISSN(Online): 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

characteristics to reduce the problem of dimensionality avoiding the repetition of words and the appearance of words with the same meaning. The increasing use of textual data requires text extraction, machine learning and methodologies to organize and extract document models and knowledge.

## REFERENCES

[1] J.-T. Chien, Hierarchical theme and topic modeling, IEEE Trans. Neural Netw.Learn.Syst.,vol.27,no.3,pp.565578,2016.
[2] Bernardini, C. Carpineto, and M. DAmico, Full-subtopic retrieval with keyphrase-based search results clustering, in IEEE/WIC/ACM Int. Joint Conf.WebIntell.IntelligentAgentTechnol.,2009,pp.206213.
[3] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, Self-organization of a massive document collection,IEEE Trans.NeuralNetw.,vol.11,no.3,pp.574585,2000.
[4] K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram, A hierarchical monothetic document clustering algorithm for summarization and browsing search results, in Proc. Int. Conf. World Wide Web, 2004, pp.658665.
[5] R. Xu and D. Wunsch, Survey of clustering algorithms, IEEE Trans. NeuralNetw.,vol.16,no.3,pp.645678,2005.
[6] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, Inductive learning algorithms and representations for text categorization, in Proc. Int. Conf. Inform.Knowl.Manag.,1998,pp.148155.
[7] R. Kohavi and G. H. John, Wrappers for feature subset selection,Artif. Intell.,vol.97,no.1,pp.273324,1997.
[8] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, Self-organization of a massive document collection,IEEE Trans.NeuralNetw.,vol.11,no.3,pp.574585,2000.
[9] Q. Mei, X. Shen, and C. Zhai, Automatic labeling of multinomial topic models, in Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2007, pp.490499.
[10] K. Lagus and S. Kaski, Keyword selection method for characterizing text document maps, in Int. Conf. Artificial Neural Networks (ICANN), 1999, pp.371376.