ISSN 2348 – 8034
Impact Factor- 3.155

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

FUZZY BASED TEXT DOCUMENTS CLUSTERING USING SIDE INFORMATION IN DATA MINING

Mr. Yuvraj R.Gurav[*1] and Assoc. Prof. P. B.Kumbharkar [2]

[*1]Research scholar, Computer Engineering, Siddhant College of Engineering Sadumbare, Pune, India

[2] Assoc. Prof. P. B.Kumbharkar, Computer Engineering, Siddhant College of Engineering Sadumbare, Pune, India

ABSTRACT

The investigation of this point is occupied with successful clustering and mining methodology with the usage of side data. The side data contained in various writings mining applications, this data may be of distinctive structures, for instance, provenance information of the records, the connections in the information, web logs that contain customer access behavior or other substance record that are embedded into the non-text based characteristics. These properties may contain a huge amount of data for clustering purposes. Be that as it may, the concerned imperativeness of this side-information may be dubious to count, especially when a rate of the information is loud. In such cases, it can be hazardous to include side-data into the mining procedure, on the grounds that it can either update the way of the representation or can include commotion in the framework, likewise partitioning algorithm is touchy against loud information Therefore, writing study proposes approach to outline effective framework that unites partitioning algorithm with fuzzy logic model for viable clustering methodology, in order to augment the benefits from the utilization of side data.

Keywords: *side information, text document clustering, feature selection, probabilistic model, fuzzy features, fuzzy logic.*

I. INTRODUCTION

The topic of text clustering considered for some application spaces, for instance, the web, casual groups, and computerized data. The rapidly extending measures of content data in sweeping online aggregations have incited an eagerness for making versatile and feasible mining algorithms. Unending effort has been done in continue going couple of years on the issue of clustering in substance gatherings in the database and information recuperation social requests. Regardless of, this work is basically expected for issue of immaculate text clustering without considering distinctive sorts of side information. A couple of tests of such side-information are according to the accompanying [1].

- We got the web logs in an application that contains Meta information related to skimming behavior of diverse customers, this kind of information can be used to upgrade the way of the substance mining, in light of the way that the logs can routinely handle fine relations in substance, that can't be understood by the foul substance alone.
- Different content archives were having lots of links inside them these links contains gigantic measure of data that is important for mining. These links go about as characteristics and as in past case these traits gives perceptions about the relations among archives in a form that may not be effortlessly conceivable from crude data.
- Meta-data associated with various web documents contrast with different sorts of characteristics, for instance, the provenance or other information in diverse cases, data, for instance, domain, position, or even common Information may be educational for mining purposes.

While such side-information can on occasion be gainful in redesigning the way of the clustering procedure, it can be a risky when the side-information is boisterous. In such cases, it can truly hurt the way of the mining philosophy. Thus, we will use a philosophy that deliberately decides the insight of the clustering characteristics of the side information with that of the text documents.

The essential thought from the examination of paper presented by Charu C. Aggarwal [2] is that, to intercede a grouping in that the substance properties and side-information passes on near sign about the method for the fundamental groups, and meanwhile reject those edges in which incongruent clues are given. To acquire this, we will combine a partitioning algorithm with a fuzzy logic system that chooses the connection of the side-qualities in the clustering process, and aides in expelling the commotion from different attributes.

ISSN 2348 – 8034
Impact Factor- 3.155

Literature survey

The issue of text-clustering has been studied with scalable clustering by the database community. [3] The issue of clustering has also been studied in the context of text-data a review of text clustering found in [4] the scatter-gather method that is the combination of agglomerative and partition clustering for text-clustering found in [5].Co-clustering methods for text data are found in [6], Matrix-factorization techniques in that words are selected from the document based on their relevance to the clustering process, and uses an iterative Expectation maximization method in order to refine the clusters [7]. A method for topic-driven clustering for text data has been proposed in [8]. Methods for text clustering in the context of keyword extraction are discussed in [9]. concept-based mining model that, analyze terms of the sentence, document, and collection level concept-based analysis algorithm concept-based similarity measure algorithm found in [10]that having limitations that this module does support for text classification and it is not linked to web document clustering. Dirichlet Process Mixture Model for Document Clustering with Feature Partition, that uses variational inference algorithm found in [11] that having limitations of adaption with the semisupervised document clustering and utilization of extra information. Clustering research proposals according to their similarities in research area found in [12] that is ontology-based text-mining approach requires extra work to cluster external reviewers depends on their research field. The text clustering in the context of scalability has also been studied in [13]. However, all of these methods are designed for the case of real text data and do not consider when text-data is combined with other forms of data. In clustering text some limited work has been done for network-based linkage information [14], [15], it is not useful to the side information attributes. In this paper, we will provide the approach to using other kinds of information in combination with text clustering. Such an approach is extremely useful, when the side information is hugely informative and provides valuable directions in creating more coherent clusters.

II. PROPOSED SYSTEM

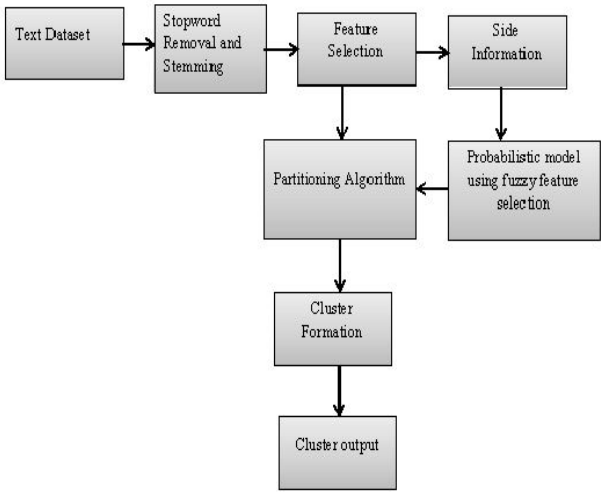

Figure 1: System Architecture

1. Text Dataset

i) 20 Newsgroup: this dataset comprise of 20000 messages taken from 20 newsgroups, this news gathering is put away in subdirectory, with each one article put away as a different document.

ii) DBLP Dataset: The DBLP information set that contains four information mining related examination ranges that are DB, information mining, IR and machine learning. This information set contains number of creators, and the writings connected with the paper distributed by these writers. This information set additionally contains the data about creator's distributions in gathering, and in that creator meeting data co-initiation considered as another sort of side data

2. Stopwords Removal

In figuring stop words can't avoid being words that are sifted out before or in the wake of transforming of common dialect information. Stop words are normal words that convey less vital significance than catchphrase. These stop

ISSN 2348 – 8034
Impact Factor- 3.155

words are without a doubt the most general, short limit words, for instance, the, is, at and so forth. Stop-word removing algorithm evacuates these words that enhance content mining execution.

Stemming

Stemming is the methodology of expelling prefix and additions from word to decrease it to stem. Stemming all around dismisses the prefix and increases they are as ing, s, es et cetera. Stemming projects are normally alluded to as stemming algorithms or stemmers. The algorithm utilized for stemming is Porter stemmer

3. Feature selection

In this procedure loud traits that are not identified with the class name are expelled from the content information and the side data separated from side characteristics. Side properties may be connections in archive, provenance of record, web logs, and other non-printed characteristics presents into the content document.as well as supplementary data in the reports. Feature determination for single-term similitude measure is a capacity contains taking after components, the cosine similarity measure is utilized alongside the Term Frequency/Inverse Document Frequency (TF-IDF) term weighting and record recurrence. The cosine measure is generally utilized as a part of the record clustering that ascertains the cosine of the point between the two archive vectors.

4. Partitioning algorithm

For dividing content information we are utilizing K- means algorithm it utilizes an arrangement of k seed centroids, around that the bunches are fabricated[17]. K-means utilizes the idea of a centroid that is the mean of a troop of focuses. The easiest manifestation of the K-means methodology is to begin with an arrangement of k seeds from the first accumulation, and appoint records to these seeds on the premise of nearest comparability. After every emphasis the seed centroid is supplanted to new seed to accomplish better essential issue for each group. This methodology is proceeded until meeting.

i) K-means algorithm

Input: S (instance set), K (number of clusters)

Output: clusters

1. Initialize K clusters centers.
2 while termination condition is not satisfies do
3 Assign instance to closest cluster center.
4.Update cluster centers based on assignments.
5 end while

ii)Algorithm complexity

Complexity of T iteration of K- means algorithm performed on sample size of m instance each characterized by N attributes, is O(T* K* m* N) this is linear complexity is one of the reason for popularity of the K-means algorithm.

5. Side Information

Here the Side or Auxiliary data is info, such side data is may be of distinctive sorts, for example, provenance of archive, the connections in information, web logs, and other non-printed qualities presents into the content record.

6. Probabilistic Model

In every Auxiliary stage, we make a probabilistic model, which in view of fuzzy feature. The partitioning algorithm is touchy against boisterous information so we utilize fuzzy feature that serves to maintain a strategic distance from the loud information and give likely related information to particular group. The point of this demonstrating is to expel uproarious information from side-data. To attain to this likelihood following fuzzy feature are considered.

7. Fuzzy Feature selection

1) Sentence Features-

In auxiliary phase, every sentence of the archive is spoken to as characteristic vector of features. In the assignment of sentence determination these highlights go about as characteristics [16]. We concentrate on a few features for every sentence that serves to expel boisterous information from side data.

ISSN 2348 – 8034
Impact Factor- 3.155

i) Sentence length:
The degree of the amount of words appearing in the sentence over the amount of words appearing in the longest sentence of the report.

$$F1 = \frac{\text{No. of the word in the sentence}}{\text{No. of words in the longest sentence}}$$

ii) Sentence Position
The initial 5 sentences in a section has a score estimation of 5/5 for first sentence, the second sentence has a score 4/5, etc.
$F2(S\,1) = n/n$; $F2(S\,2) = 4/5$; $F2(S\,3) = 3/5$ $F2(S\,4) = 2/5$ so on...

iii) Numeric Data
The level of the incorporate of numerical data sentence to the sentence length.

$$F3\,(Si) = \frac{\text{No. of the Numerical data in the sentence Si}}{\text{Sentence length}}$$

iv) Term Weight

It is the proportion of summation of term frequencies of all terms in a sentence over the greatest of summation estimations of all sentences in a report.

$$F4 = \frac{\Sigma TFi}{MAX\,(\Sigma TFi)}$$

Where TF=Term Frequency, i =1 to n, n is the number of terms in a sentence.

v) Sentence to sentence similarity
For every sentence S, the comparability in the middle of S and each other sentence is figured by the technique for token matching.

$$F5 = \frac{\Sigma\,[Sim(Si,Sj)]}{MAX\,[Sim(Si,Sj)]}$$

Fuzzy Logic-

FL is used for solving uncertainties in a given problem. Fuzzy logic system design usually consists of selecting fuzzy rules and association role. The fuzzy logic system consists of following components fuzzier, inference engine, defuzzifier, and the fuzzy knowledge base. In the fuzzifier, crisp inputs are converting into grammatical codes using association role to be used to the input grammatical variables. After fuzzification, the inference engine refers to fuzzy IF- THEN rules to derive the grammatical codes. In the last step, the output grammatical variables from the inference are converted to the final crisp values by the defuzzifier using association role for representing the final score of sentence. Thus each sentence is associated with 5 features. Using all the feature scores, the score for each sentence are derived using fuzzy logic method. The fuzzy system associated with fuzzy rules and triangular association role. The fuzzy guidelines are as IF-THEN. The triangular affiliation part fuzzifies each one score into one of 3 values that is LITTLE, AVERAGE, BIG. At that point we apply fuzzy rules to figure out if sentence is immaterial, normal or important. This is also called as defuzzification.
For example IF (F1is H) and (F2 is M) and (F3 is H) and (F4 is M) and (F5is M) THEN (sentence is important).All the sentences of a document are ranked in a descending order based on their rates. Higher n sentences of highest score are taken out as document clustering based on compression rate. Finally the sentences in cluster are arranged in the order they occur in the original document.

ISSN 2348 – 8034
Impact Factor- 3.155

8. ***Cluster formation and Output cluster***

Cluster formation is stage in that all the data grouped utilizing side Information, fuzzy feature chooses imperative Data by dodging undesirable information from side data and partitioning algorithm with this fuzzy probabilistic model upgrade nature of the cluster. At long last, the proposed model gives quality clusters by exploiting side data.

III. RESULT & DISCUSSION

In this segment, we look at our clustering methodology against different datasets. We utilized the K-means in addition to fuzzy framework that gives brilliant clustering results in an exceptionally effective manner. We proposed our probabilistic model of fuzzy with administered clustering technique that uses both content and side data.

Thus, we compare our model results with standard data mining techniques precision and recall that leads advantage of our approach over both a pure text-mining method and a natural alternative that uses both text and side information.

- Precision is the fraction of retrieved documents that are relevant to the field.
- Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved

In figure 3 the accuracy comparison graph shows that the accuracy is calculated by using precision and recall method

Where we have provided 1000 records in dataset 3 among them we are expecting 200 positive related records from the initially accessed 300 records. In first access we found 60 related records now, for accuracy comparison we consider following cases

1.TN /True Negative : case was negative and predicted negative.
2. TP / True Positive : case was positive and predicted positive.
3. FN / False Negative : case was positive but predicted negative.
4. FP / False Positive: case was negative and predicted positive.
Where TP=60, FP=240, TN=860, FN=140.

The accuracy of our system for input Dataset 3 is (TN + TP) = (860+60) out of 1000 result accuracy will be up to 92%. Where the old system showing 90% accuracy that is less than our system.

Results of Practical Work

Table 1. Feature **selection score**

input	f1	f2	f3	f4	f5	Score
Dataset 1	0	1	0	1	1	0.6
Dataset 2	1	1	1	1	0	0.8
Dataset 3	0	0	1	0	0	0.2
Dataset 4	0	1	1	0	0	0.4

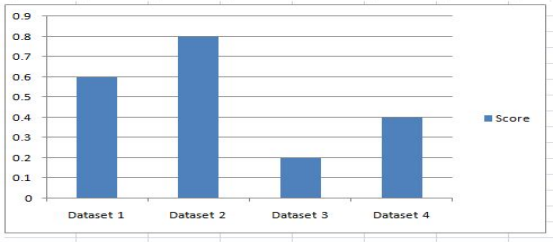

Figure2. Fuzzy feature selection score.

We selected F1, F2, F3, F4 and F5 sentence selection features to reduce noisiness from information source. fig.2 shows that average scoring for sentence selection where X axis represent source information Y axis indicate average score.

ISSN 2348 – 8034
Impact Factor- 3.155

Table2. Accuracy comparison

Comparision Graph		
Input	K-means	K-means+Fuzzy
dataset1	78	80
dataset2	70	75
dataset3	90	92
dataset4	67	70

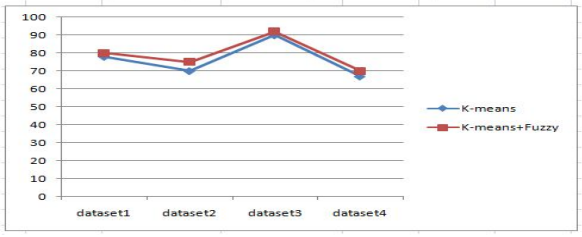

Figure3. Comparison our system with K- means

Fig.3 shows that K-means plus fuzzy give better accuracy of clustering output. Accuracy measurement reflects really the difference between our fuzzy technique and earlier one. If input data is highly inaccurate it drives evaluation of result poor else it is decent. In the graph X-axis indicate source of information and Y-axis represent accuracy in percentage.

Table3. Time complexity

Time Complexity	
Input	Output(Sec.)
dataset1	15
dataset2	10
dataset3	20
dataset4	25

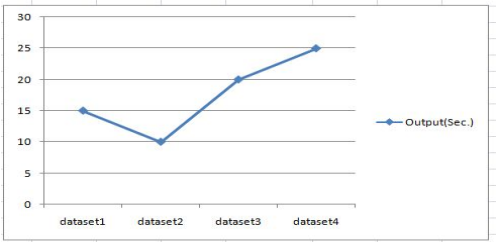

Figure 4. Time complexity

This graph shows time complexity of our system that signifies the total time required by the program to run to completion. Time complexity varies with size of dataset where X axis shows the source of information and Y-axis shows the execution time in seconds.

IV. CONCLUSION

In this paper, we displayed systems for mining text data with the usage of side or auxiliary data. Numerous manifestations of content datasets contain a lot of side data, for example, provenance of archive, the connections in information, web logs, and other non-textual traits presents into the content report that may be used to enhance nature of clustering system. With a particular finished objective to diagram the clustering framework, we united an iterative partitioning algorithm with a fuzzy probabilistic model that forms the centrality of different sorts of side information. The experimental result of our model exhibit that the use of side information can phenomenally enhance the way of clustering, while keeping up an anomalous condition of benefit. We can extend our work to enhance grouping proficiency of web information.

V. ACKNOWLEDGEMENTS

We are glad to express our sentiments of gratitude to all who rendered their valuable guidance to us. We would like to express our appreciation and thanks to Dr.S.S.Khot, Principal, Siddhant College of Engineering

ISSN 2348 – 8034
Impact Factor- 3.155

Sadumbare. We are also thankful to our family and friends for their encouragement and support. We thank the anonymous reviewers for their comments

REFERENCES

1. *Y.R.Gurav,P.B. kumbharkar "A Review on Side Information Entangling For Effective Clustering Of Text Documents in Data Mining", IJCSIT., vol. 5(6), pp. 7239-7242, Nov. 2014*
2. *Charu C. Aggarwal, Yuchen Zhao, and Philip S. Yu, "On the Use of Side Information for Mining Text Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 6,pp 1415-1429, June 2014*
3. *S. Guha, R. Rastogi, and K. Shim, "An efficient clustering algorithm for large databases", ACM ,NY USA,pp 73-84,1998*
4. *C. C. Aggarwal and C.-X. Zhai," Mining Text Data", Springer, NY, USA, 2012.*
5. *D. Cutting, D. Karger, J. Pedersen, and J. Tukey, "Scatter/Gather: A cluster-based approach to browsing large document collections", ACM SIGIR Conf., NY, USA, pp.318–329, 1992*
6. *I. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning", ACM KDD Conf., NY, USA, pp. 269–274 2001.*
7. *W. Xu, X. Liu, and Y. Gong, "Document clustering based on nonnegative matrix factorization", ACM SIGIR Conf., NY, USA, pp. 267–273, 2003.*
8. *Y. Zhao and G. Karypis "Topic-driven clustering for document datasets", SIAM Conf. pp. 358–369, 2005.*
9. *H. Frigui and O. Nasraoui, "Simultaneous clustering and dynamic eyword weighting for text documents",Springer, NY, USA, pp. 45–70, 2004.*
10. *Shady Shehata, Fakhri Karray"An Efficient Concept based Mining Model for enhancing text clustering", IEEE transaction on knowledge and data engineering, vol.22, no.10, pp.1360-1371, 2010.*
11. *]Ruizhang Huang, Guan Yu, Zhaojun Wang, Jun Zhang, and Liangxing Shi " Dirichlet Process Mixture Model for Document Clustering with Feature Partition", IEEE Transactions On Knowledge And Data Engineering, vol. 25, no. 8, pp.1748-1759,2013*
12. *Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang,and Ou Liu " An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection", IEEE Transactions On Systems, Man, And Cybernetics, vol. 42,no. 3, pp.784-790, 2012*
13. *C. C. Aggarwal and P. S. Yu, "A framework for clustering massive text and categorical data streams", SIAM Conf. Data Mining, pp. 477–481, 2006.*
14. *C.C.Agrawal and H.Wang,"Managing and Mining Graph Data", Springer, NY, USA, 2010.*
15. *C. C. Aggarwal, "Social Network Data Analytics", Springer, NY, USA, 2011.*
16. *Ladda Suanmali ,Naomie Salim and Mohammed Salem Binwahla "Feature-Based Sentence Extraction Using Fuzzy Inference rules", IEEE,978-0-7695-3654-5 ,2009*
17. *H. Schutze and Silverstein, Projections for Efficient Document clustering.in Proc. ACM SIGIR Conf., NY,USA, pp 7481,1997.*