International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Efficient Text Mining Model with Conceptual Informative Relational Measure using Semantic Ontology

G Shobarani, K Arulanandham

Abstract: ***The problem of text mining has been well studied and numerous approaches are analyzed towards their performance in text mining. The existing methods suffer to achieve higher performance as they consider only content of document and the term features available. Also, they measure the similarity between documents on the term features to identify the class of any document. This affects the performance of text mining and produces poor accuracy and generates higher irrelevancy. To improve the performance, a Conceptual Informative Relational Model (CIRM) is presented in this paper. Unlike previous methods, the method considers both conceptual and informative relations in measuring the similarity between the documents. The method preprocesses the text documents by eliminating the stop words, stemming and identifies list of root words or nouns. The root words extracted has been used to measure the conceptual relation and informative relation according to the taxonomy of classes and semantic meanings. Based on the value of relational measures, the method identifies the class of the document and produces result set. The proposed method improves the performance of text mining and reduces the irrelevancy.***

Index Terms:Text Mining, Semantic Ontology, CIRM, Relation, CRM, IRM, CISM.

I. INTRODUCTION

The growth of information technology has allowed the organization to maintain various information in different forms. The organization would maintain different information in form of documents. There would be number of documents available in any organization and they will be related to different concepts. As the number of documents increases, retrieving required document from the large pool of documents becomes more challenging task. So it is necessary to organize them in different classes. However, retrieving the related document becomes another issue, which is performed by several techniques.

Text mining is the process of extracting related documents from huge set of documents. The retrieval of documents is performed according to the similarity between the documents. The similarity between the documents is measured in several ways. The K-means algorithm measure the document similarity based on the distance between the points or terms. Similarly, the term frequency and inverse document frequency based algorithm measures the similarity based on TF-IDF.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
G Shobarani*, Ph.D Research Scholar, Bharathiar University, Coimbatore, Tamilnadu, India
K Arulanandham, Head, Department of Computer Applications, Government Thirumagal Mills College, Gudiyattam, Tamilnadu, India

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

Similarly, there are number of approaches available for the problem of text mining. All the methods consider the text features present in the document. According to the text features of the document, the methods measure the similarity between the query and document. So, the most methods consider only the textual features and they produce higher false ratio and irrelevant results.

Further, the category of the document would be classified into number of levels. For example, the category "Computer" can be classified into Computers, Computers / Programming Language, Computers / Programming Language / Java. Similarly, the categories of the document can be classified up to any number of levels. The text mining algorithm should consider the exact category and the sub classes of the data set. However, different features are used for text mining, considering the topical and informatics measures are more important.

By considering the text features in text mining, the methods only count the presence of terms in the document or class of documents. But they do not consider the concept need to be discussed and how well the topic is discussed. According to this there are algorithms which consider the conceptual measures and they also suffer to produce efficient results. However, the method does not consider the relationship between the documents in terms of concept and semantic meanings. In order to become a member of the class, the document has to discuss the concept at the maximum level and the terms of the class should be well discussed. According to this, the author has been motivated to design an efficient algorithm for text mining.

Semantic ontology is the taxonomy which has number of classes and features where there will be number of relations presented between the features and concepts. So by considering the relations between the conceptual terms the conceptual relation of the document can be measured. Similarly, the relation on informatics terms can be used to measure the semantic relation of any query and document. By considering both conceptual and semantic relations of the document, the document similarity can be measured. The detailed approach is discussed in the next section.

II. RELATED WORKS

There are number of techniques discussed towards text mining in literature. This section discusses a set of methods related to the problem. In [1], the author presents various text mining techniques and applications in detail.

The method present different discovery patterns to analyze documents from huge volume of data set. The text mining process has been discussed as patterns and features based algorithm in extracting related documents.

Using text mining to classify research papers, [2] uses natural language processing tools towards the classification of research papers. The method has been adapted support vector machine and naïve bayes algorithms in classification.

In [3], the authors present a trend recognition algorithm for journal papers. The popular TF and IDF algorithms have been used. The TF measure fetches the topical strength of the document where IDF measure fetch the topical strength of other categories. The method has been adapted in quality control of Japanese documents.

In [4], the Korean deciphering monetary board data has been mined using field specific keywords to support the communication of central bank of korea. The method fetches the indicators based on text to support the explanation of BOK monetary decisions. The method produces efficient results on textual classification and media based measure to support economic policy.

In [5], a sentimental analysis approach to support social media has been presented. The method analyzes the social media data sets by constructing a dictionary of words. Based on the words of dictionary, the method selects a set of tags on any topic. Based on the polarity of tags, the method performs classification of tweets. The US election data set has been used for evaluation.

Text-based predictions of beer preferences by mining online reviews [6], discusses the process involves decomposing the texts into the words composing them and using the frequency of those words to predict the text polarity. The text categorization approach may use single words composing the texts or longer combinations of the words.

Longer combinations of the words have the advantage of better representing the complexity of human language compared to single words. Moreover, less frequent terms may also better discriminate between reviews than more common words.

This study tests these two hypotheses in the context of beer reviews. It shows that the words used in the reviews can be used to predict consumer's preferences for beer. Moreover, it shows that the use of less frequent terms in the predictive models outperforms the use of more frequent terms.

In [7], the author presented an opinion classification algorithm towards the development of tourist reviews. Earlier systems used implicit, infrequent measures in classifying the opinion. In earlier approach, an aspect based classification is presented which suffers with higher irrelevancy.

To improve the performance, a fuzzy based algorithm is presented which extract the aspects according to the opinions of user. The method produces accurate classifications than previous algorithms.

In [8], a sentiment analysis algorithm is presented which uses lexicon to analyze the performance of teachers. The method supports the evaluation of teacher's performance by obtaining feedback from students. From the feedback obtained from students, the text mining techniques have been used to analyze the feedback to promote performance evaluation of students.

In [9], the author describes the idea of converting the Ecommerce data into classification problem. The method applies text mining algorithms to identify the similar records of e-commerce data.

The method verifies six different strategies towards the selection of key words and grams. Finally, the classification is performed with SVM, Random Forest, and Logistic Classifier.

In [10], the authors present a prospective analysis based on the study of PubMed search history enables us to determine the possible directions for future research.

In [11], the authors present a performance analysis algorithm for Airline marketing data using text mining approaches. The method first extracts the key words from the market data and identifies the most prominent terms. Further the method estimates the influence of key word towards corporate performance analysis.

In [12], a semantic pattern mining algorithm has been presented towards text mining. The method generates number of semantic pattern mining and measures the frequency of semantic patterns. Such pattern frequency has been sorted using suffix array sorting to perform semantic pattern mining.

In [13], a user interest prediction model is presented which combines Latent Dirichlet Allocation (CLDA), which uses big data. The method identifies and learns the topics from blogs of big data. The method learns the short text from long text towards the improvement of text mining.

In [14], the authors presents a bug report classification algorithm using text mining. The method uses text mining algorithm in interest prediction in multiple stage approach. First the method analyzes the text features from bug reports and estimates the probability to classify them in three levels. The learned features are applied with machine learning to perform prediction.

In [15], a sentiment analysis of reviews towards the scholarly papers has been presented. The method automatically predicts the recommendation generated for any article by analyzing the statements of reviews. The method estimates the polarities to perform classification.

Towards the development, a multiple instance learning network with abstract based memory mechanism (MILAM) is presented.

In [16], the authors review various approaches of web document clustering. The paper discusses various aspects and algorithm which has been used for mining text from web documents. Finally, a harmony search algorithm has been presented towards the grouping of web documents. Different algorithms have been compared for their performance in different parameters.

All the above discussed methods suffer to achieve higher performance in text mining and document classification.

III. CONCEPTUAL INFORMATIVE RELATIONAL MEASURE BASED TEXT MINING (CIRM)

The proposed algorithm takes the input data set and for each document the text features has been extracted to perform preprocessing which eliminates stop words, stemming and identifies list of root words or nouns. The root words extracted has been used to measure the conceptual relation and informative relation according to the taxonomy of classes and semantic meanings. Based on the value of relational measures, the method identifies the class of the document and produces result set. The detailed algorithm is presented in this section.

The Figure 1 shows the architecture of proposed conceptual and informative relational model for efficient text mining. Each stage has been described in detail.

A. Preprocessing:

The preprocessing is performed on the documents given and from each document the text features has been extracted. From the text features extracted, the method identifies the stop words which has no meaning and does not support any way. The remaining textual terms are stemmed to identify the root words. Identified root words are analyzed for their importance by applying part of speech tagger provided by Stanford University. Based on the result of tagging, a list of nouns has been identified as keywords to perform text mining.

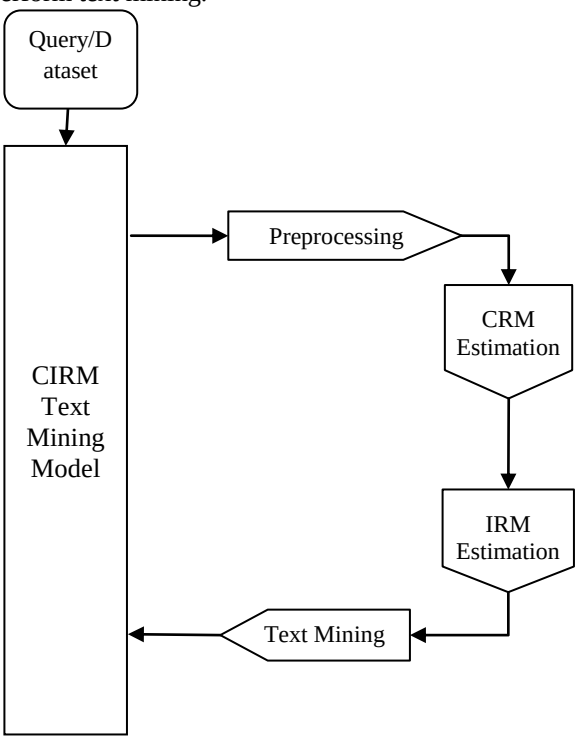

Figure 1: Architecture of Proposed CIRM Text Mining Model

Algorithm:

Input: Document D, Stop words Sw, Tagger Pos
Output: Term Set Ts
Start

Read document and extract text features T.
Identify the term set Ts= $\int \text{Split}(T,)$
Eliminate stop words based on stop word list Sw.

$$Ts = \int_{i=1}^{size(Ts)} \text{if } Ts(i) \in Sw \text{ then } Ts \cap Ts(i)$$

Now stem the words to get Root words.

$$Ts = \int_{i=1}^{Size(Ts)} Ts(i) \cap \{ing, ed\}$$

Now Tag the words to identify pure nouns.

$$Ts = \int_{i=1}^{Size(Ts)} \text{if } Pos(Ts(i))! = Noun \text{ then } Ts = Ts \cap Ts(i)$$

Stop

The above discussed algorithm shows how the preprocessing of text documents are performed to get the pure nouns from the document text. The method reads the input text from document and eliminates the stop words and stems the words to get root words. Finally, the words are tagged to get the pure nouns to perform text mining.

B. Conceptual Relational Measure (CRM) Estimation:

The conceptual relation of any document has been measured according to the frequency of terms in the concept taxonomy. The taxonomy maintains number of terms in hierarchical form. Based on the concept terms from the taxonomy and the terms of document being extracted, the method measures the conceptual relational measure. First, the method finds the number of terms of terms set extracted which are appear on the concept taxonomy. Second according to the number of terms of taxonomy and the appearance measure, the method computes the CRM measure.

Algorithm:

Input: Taxonomy T, Term Set Ts
Output: CRM.
Start

Read term set Ts and Taxonomy T.
Initialize Count to 0.
For each term Ti from Ts
If

$$\int_{i=1}^{size(T)} \text{if } T(i) == Ti \text{ then } count++$$

End

Compute CRM = $\frac{Count}{size(T)}$

Stop

The principle of estimating conceptual relational measure towards any concept or category has been presented in the above discussed algorithm. The frequency of terms of corpus has been measured and based on the total terms of the concept, the method estimates the CRM measure.

C. Informative Relational Measure (IRM) Estimation:

The informative relational measure represents the fitness of document in presenting more information related to any category. To become more informatics it is not necessary to discuss only the terms related to a category. To measure this, the wordnet dictionary has been used which has been used to generate the ontology for any category.

For each terms of term set extracted in the preprocessing stage, the method extract the related terms from wordnet to frame the ontology of any category. Based on the semantic terms of category and the terms of term set, the method estimates the informative relational measure towards any specific category.

Algorithm:

Input: Term Set Ts, Word Net Wn
Output: IRM
Start
Read Ts, T, Wn.
Initialize Ontology O.
For each term Ti from Ts
Extract relational terms from word net.

$$O = \int_{i=1}^{size(T)} O \cup (\sum Wn(Ti.Synonyms) \in Wn)$$

End

Compute IRM = $\frac{\int_{i=1}^{size(Ts)} \sum Ts(i) \in O}{size(O)}$

Stop

The informative relation measure on specific document towards any category has been measured based on the synonyms obtained from word net. The working principle of estimating informative relational measure has been presented in the above discussed algorithm.

D. CIRM Text Mining:

The conceptual and informative relational similarity measure based text mining algorithm starts with preprocessing the input query and extracts the terms to measure the conceptual similarity measure. Based on the value of CRM being measured, the method identifies the class with higher conceptual relation. Second, the document of class has been extracted and for each of them, the conceptual relational measure and informative relational measure for each document has been measured. Using these two values, the method estimates the CIRM weight for each of them. According to the value of CIRW, the method selects a subset of documents to produce result.

CIRM Text Mining Algorithm:

Input: Query Q, Data Set Ds, Taxonomy T, Ontology O
Output: Result Set Rs
Start
Read query Q.
Term set Ts = Preprocessing (Q)
For each class c
Compute CRM measure QCRM =

$$\int_{i=1}^{size(C)} CRMEstimation(c, Ts)$$

End
Choose the class with higher CRM value.

Class C= $\int_{i=1}^{size(C)} Max(C(CRM))$

For each document Di of Ds
Ts = Preprocessing(Ts)
CRM = Estimate-CRM(Ts)
IRM = Estimate-IRM(Ts)
Compute CIRW = CRM×IRM
If CIRW>Th Then
Add Di to Result set.
Rs = $\sum(Documents \in Rs) \cup Di$
End
End
Stop

The conceptual and informative relational measure based text mining algorithm has been described above which estimates CRM and IRM measures. Based on these measures, the method compute the CIRW value, which has been used to select subset of documents from the class identified.

IV. RESULTS AND DISCUSSION

The proposed conceptual and informative relational measure based text mining algorithm has been implemented and evaluated for its performance in various parameters. The method has been implemented using Advanced java and the performance of the method has been evaluated under different parameters. This section present the results obtained by the proposed algorithm and present the comparative results.

Parameter	value
Tool Used	Advanced Java
Data Set	Reuters, KDD, UCI
Number of Class	15
Number of Documents	10 lakh

Table 1: Evaluation Details

The Table 1 shows the details of evaluation being used to measure the performance of various methods of text mining. The performance of the algorithms has been measured under the following parameters.

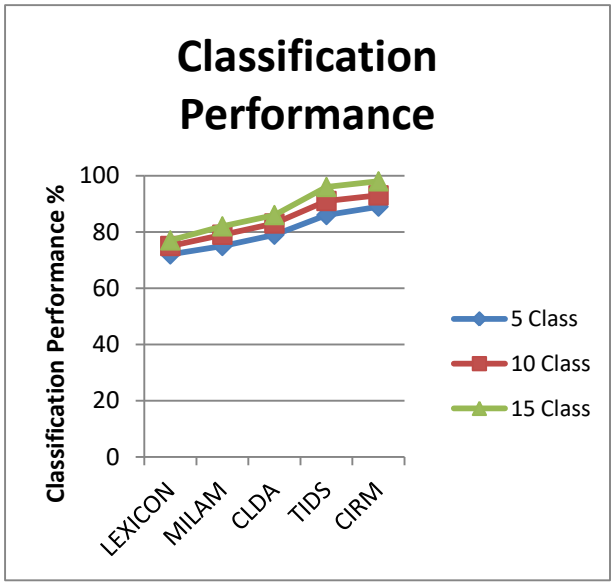

Figure 2: Performance on Classification

The performance on classification accuracy has been measured for different algorithms. The result produced has been compared and presented in Figure 2.

The proposed CIRM text mining algorithm has achieved higher classification performance than other methods.

The text mining performance of different methods has been measured and compared with the result of proposed methods. The proposed CIRM Text mining algorithm has produced higher text mining performance than other methods. The performance analysis result is presented in Figure 3.

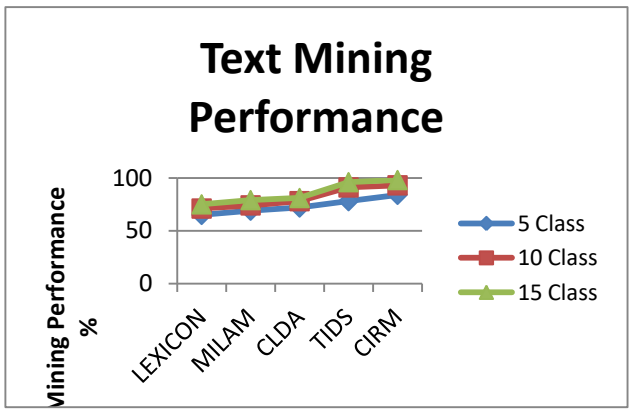

Figure 3: Performance on text mining

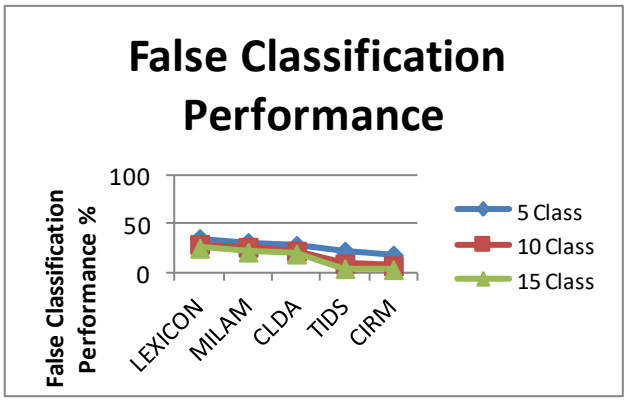

Figure 4: Performance on False classification ratio

The false classification ratio has been measured for the proposed CIRM text mining algorithm and compared with the result of other methods. The proposed CIRM algorithm has reduced the false ratio than other methods.

The performance on time complexity has been measured and compared with the results of other methods and presented in Figure 5. The proposed CIRM algorithm has achieved less time complexity than other methods.

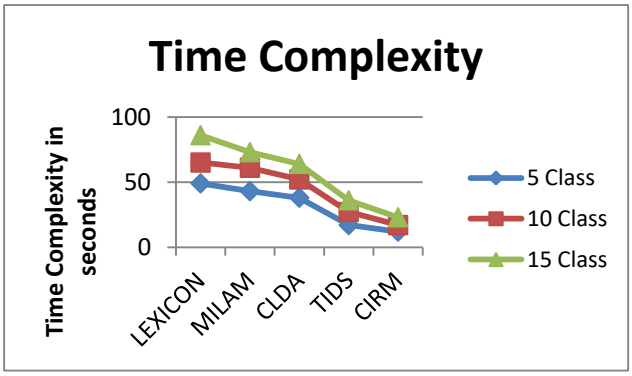

Figure 5: Performance on Time Complexity

V. CONCLUSION

The problem of text mining has been approached with an efficient conceptual and informative relational measure based algorithm. The query has been preprocessed to get the nouns and measure conceptual relational measure towards various classes. Based on the value of CRM, a single class with higher CRM value has been selected. From the selected class, the method preprocesses each document and estimates CRM and IRM values for each class. Using the value of CRM and IRM value, the CIRM weight is measured. According to the value of CIRM weight, the method selects a sub set of documents according to the threshold. The selected documents are populated as result to the user. The proposed CIRM algorithm has achieved higher performance in text mining and classification. The false ratio and time complexity has been hugely reduced.

REFERENCES

1. RamzanTalib, Text Mining: Techniques, Applications and Issues, (IJACSA), Volume 7, Number 11, 2016.
2. Sulova, Snezhana, Using text mining to classify research papers, (SGEM), Volume 17, 2017.
3. M. Terachi, R. Saga and H. Tsuji, Trends Recognition in Journal Papers by Text Mining, (IEEE) 2006, pp. 4784-4789.
4. Park, Ki Young, Deciphering Monetary Policy Board Minutes Through Text Mining Approach: The Case of Korea", (SSRN), 2019.
5. Imane El Alaoui, A novel adaptable approach for sentiment analysis on big social data, (Springer, Journal of Big Data), 2018.
6. AbdelazizLawani, Text-based predictions of beer preferences by mining online reviews, (IBM), 2019.
7. Muhammad Afzaal, Fuzzy Aspect Based Opinion Classification System for Mining Tourist Reviews, (HINDAWI), 2016.
8. Quratulain Rajput, SajjadHaider, and SayeedGhani, Lexicon-Based Sentiment Analysis of Teachers' Evaluation, (HNDAWI), 2016.
9. Gianpiero Bianchi, Renato Bruni, Identifying e-Commerce in Enterprises by means of Text Mining and Classification Algorithms, Mathematical Problems in Engineering, (HINDAWI), 2018.
10. Ekaterina Ilgisonis, AndreyLisitsa, Creation of Individual Scientific Concept-Centered Semantic Maps Based on Automated Text-Mining Analysis of PubMed, Advances in Bioinformatics (HINDAWI), 2018.
11. Jae-Won Hong,Seung-Bae Park, The Identification of Marketing Performance Using Text Mining of Airline Review Data, (Mobile Information Systems, HINDAWI), 2019.
12. X. Song, X. Wang and X. Hu, "Semantic pattern mining for text mining," IEEE(Big Data), 2016, pp. 150-155.
13. LirongQiu, Jia Yu, CLDA: An Effective Topic Model for Mining User Interest Preference under Big Data Background, (HINDAWI), 2018.
14. Yu Zhou, Yanxiang Tong, Combining text mining and data mining for bug report classification, (SEP), Volume28, Issue3, 2016, pp 150-176.
15. Ke Wang, Xiaojun Wan, Sentiment Analysis of Peer Review Texts for Scholarly Papers, (SIGIR), 2018, pp 175-184.
16. Forsati R., Mahdavi M. Web Text Mining Using Harmony Search, Studies in Computational Intelligence, vol 270, 2010.

AUTHORS PROFILE

Ms. G. Shobarani studied Bachelor of Science and Master of Computer Applications from University of Madras and Master of Philosophy from Bharathidasan University. Her main research interest is in the area of Data Mining, Data Warehousing, Machine Learning and Programming Languages. She has around 18 years of teaching experience and 11 years of research experience.

Dr. K. Arulanandam is currently working as Assistant Professor & Head in the Department of Computer Science and Applications, Government Thirumagal Mills College, Gudiyattam. He has published several papers in reputed National and International journals. He has 17 years of experience in teaching and research.