ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

Enriching Text Clustering through Cognitive Concept Based Model (CCBM) by Ensemble Neural Learning Method

K.Sangeetha[1], M.V. Sayini[2]

Assistant Professor, Department of PG Computer Science, PR Engineering College, Vallam, Thanjavur, India[1]

M.Phil Scholar, Department of Computer Science, PRIST University, Thanjavur, India[2]

ABSTRACT: Nowadays the text mining has become one of the emerging research technology which has been incorporated in several research fields such as computational linguistics, Information Retrieval (IR) and data mining. Text mining reads structured as well as unstructured data to provide meaningful information patterns using Natural Language Processing (NLP) techniques to extract knowledge from the textual text. The communications of social networking sites are without correct grammar and spelling which may lead to different kinds of ambiguities in lexical, syntactic, and semantic levels and so it is hard to find out the actual concept of the text. This paper focused on analyzing the concept based text mining model related social media contents.

Ensemble Neural learning method is a paradigm where multiple neural networks are jointly used to solve a problem. In this paper, the concept space relationship and its sentence, document, and corpus component neural networks is analyzed from context of classification. This leads optimum level of results through ensemble multiple neural learning networks for prediction. By combining the traditional similarity feature space and the concept extension space, the adverse effects of the complexity and diversity of natural language can be addressed and clustering can be improved correspondingly. The generated clusters can be organized using different granularities. The experimental evaluations on datasets have verified the effectiveness of our approach. Experimental result enhances text clustering quality by using sentence, document, corpus and combined approach of concept analysis.

KEYWORDS: Text mining, text clustering, Cognitive Concept Based Model, Neural Learning

I. INTRODUCTION

The text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. The resources of unstructured and semi structured information include the world wide web, governmental electronic repositories, news articles, biological databases, chat rooms, digital libraries, online forums, electronic mail and blog repositories [1]. Therefore, proper classification and knowledge discovery from these resources is an important area for research. Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the electronic documents [2].

The document classifications adapted by unsupervised, supervised and semi supervised methodologies where many techniques and algorithms are proposed for the clustering and classification of electronic documents. However, how these documents can be properly annotated and classified? [3]. So it consists of several challenges, like proper annotation to the documents, appropriate document representation, dimensionality reduction to handle algorithmic issues, and an appropriate classifier function to obtain good generalization and avoid over-fitting [3] [4]. Extraction, Integration and classification of electronic documents from different sources and knowledge discovery from these documents are important for the research communities [7].

From last few years, the task of text classification have been extensively studied and rapid progress seems in this area, including the approaches such as Bayesian classifier, Decision Tree, K-nearest neighbor(KNN), Support Vector

DOI: 10.15680/IJIRCCE.2017. 0508005

Machines(SVMs), Neural Networks, Latent Semantic Analysis, Rocchio's Algorithm, Fuzzy Correlation and Genetic Algorithms etc [2] [3].

In general text mining techniques compute the term frequency of the terms in the whole document to find the importance of the term, where the meaning contributed by one term is more suitable to the sentence meaning than the meaning contributed by the other term [7]. Hence, we need the new required model with capturing semantic structure of each term at various concepts levels rather than the term frequency of the document. Concept-based similarity determines the similarity outcomes of the concept analysis on the sentence, document and corpus levels [9].

Each sentence in a document is labeled by a semantic role labeler who determines the terms semantics contribution to the sentence. The semantic role of a term in a sentence or statements is called as concept that may be words or phrases. Our proposed cognitive concept-based similarity measure outperforms other similarity measures on the sentence, document and the corpus levels. The quality of clusters produced is influenced by the similarity measure used as it is insensitive to noise while calculating the similarity. This is because the concepts are analyzed in the sentence, document and corpus levels and hence the probability to find a concept match between unrelated documents is very small.

The rest of the paper is organized as follows. In Section 2 describes an overview of the background concepts and related works, in Section 3 presents the preliminary framework for our models, in Section 4, the experimental analysis were presented, and finally in Section 5 consists the conclusion.

II. RELATED WORK

A. THE BACKGROUND

In the booming research field of Text Mining, the meaningful information framed from language text through the process of analyzing text for particular purposes [3] [2]. The steps involved in the overall process of the text mining are depicted in the Figure 1.

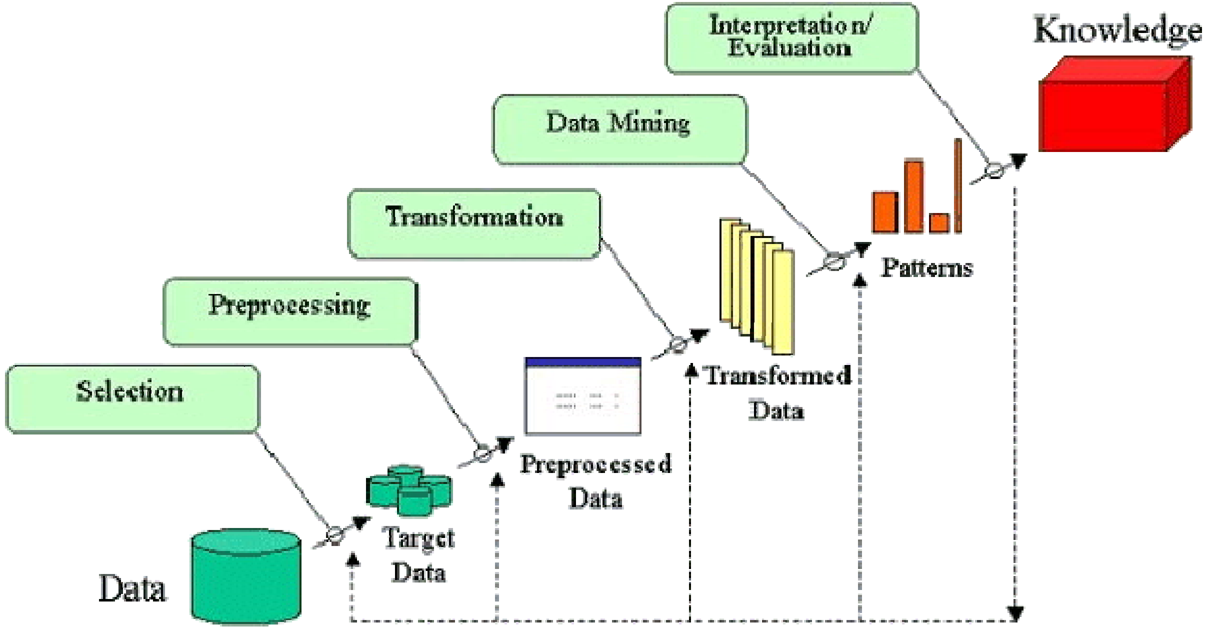


Figure 1.General text mining process flow

In clustering, organizing un-clustered text documents into groups or clusters with high intra-cluster similarity and low inter-cluster similarity [3] is most important. In general, text clustering methods try to separate the documents into groups in which each group represents a topic that is different from the other groups.

Decision trees, clustering based on data summarization, rule-based systems, statistical analysis, and neural nets are some of the methods that are used for text clustering [3]. The most important aspect in text mining is that the output of the clustering algorithm depends on the features that have been selected [1] [7]. Furthermore, the result of the clustering algorithm is based on the weights of the selected features.

B. THE REVIEW OF LITERATURE

In [8], Fillmore suggested that thematic roles are categories which help characterize the verb arguments by providing a shallow semantic language. Now the thematic roles in sentences adapted in the form of automatic labeling.

Gildea and Jurafsky [9], applied a statistical learning technique to the FrameNet Database which determining the probable role of the constituent given the predicator, frame and certain other features. These probabilities were trained on the FrameNet database through shallow semantic parsing with Support Vector Machines (SVM). They depend on the verb, the voice of the verb, the grammatical functions and other such features. Support Vector Machines are used to identify the arguments of a given verb in a sentence and also to classify them by the semantic roles that they play like AGENT, THEME, and GOAL.

The Vector Space Model [11],[10] is widely used document clustering method and represents data for text classification and clustering. The terms in the document is represented as a feature vector. The terms can be words or phrases. Each feature vector is assigned a term weight based on the term frequency of the terms in the documents. Similarity measures that rely on the feature vector is used to find the similarity between the documents (Cosine measures and the Jaccard measure).

According to the PropBank notations [12], each of the sentences in the document is labeled automatically. The sentences in the document may have one or more labeled verb argument structures. The amount of information in the sentence influences the number of verb argument structures generated by the labeller. The output of the labeller and the labeled verb argument structures are captured and analyzed by the concept based analysis model on the sentence, document and corpus levels.

NLP is one of the hot topics that concerns about the interrelation among the huge amount of unstructured text on social media [13], besides the analysis and interpretation of human-being languages [14], [15]. In general, the data related to social media sites is not collected for the research purpose [16], it is mandatory to change the structure of the data coming from the social media. 80% of the available text on the web is unstructured while only 20% is structured [17].

Nowadays, Facebook is one of the most popular social media. This media is used by a large number of people on earth for expressing their ideas, thoughts, sorrows, pleasures and poems [18]. Researchers had chosen a number of Facebook variables that were expected to develop the right situation for carrying out our investigations. The valuable statistics of user’s personality is provided by the Facebook profiles and activities, which exposes the actual objects instead of projected or idealized character [19].

Neural network learning method is a paradigm where a collection of a finite number of neural networks is trained for the same task [23] which shows that the generalization ability and they can be significantly improved through ensembling a number of neural networks, i.e., training many neural networks and then combining their predictions.

III. PRELIMINARY FRAMEWORK

Most of the clustering techniques are based on statistical analysis or any simple formulated analysis in the related work. In our system we are going to use social media documents and the clustered output using the cognitive concept based model through neural learning methodology which consider all levels of concept analysis, and concept-based similarity measure, as depicted in Figure 2.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

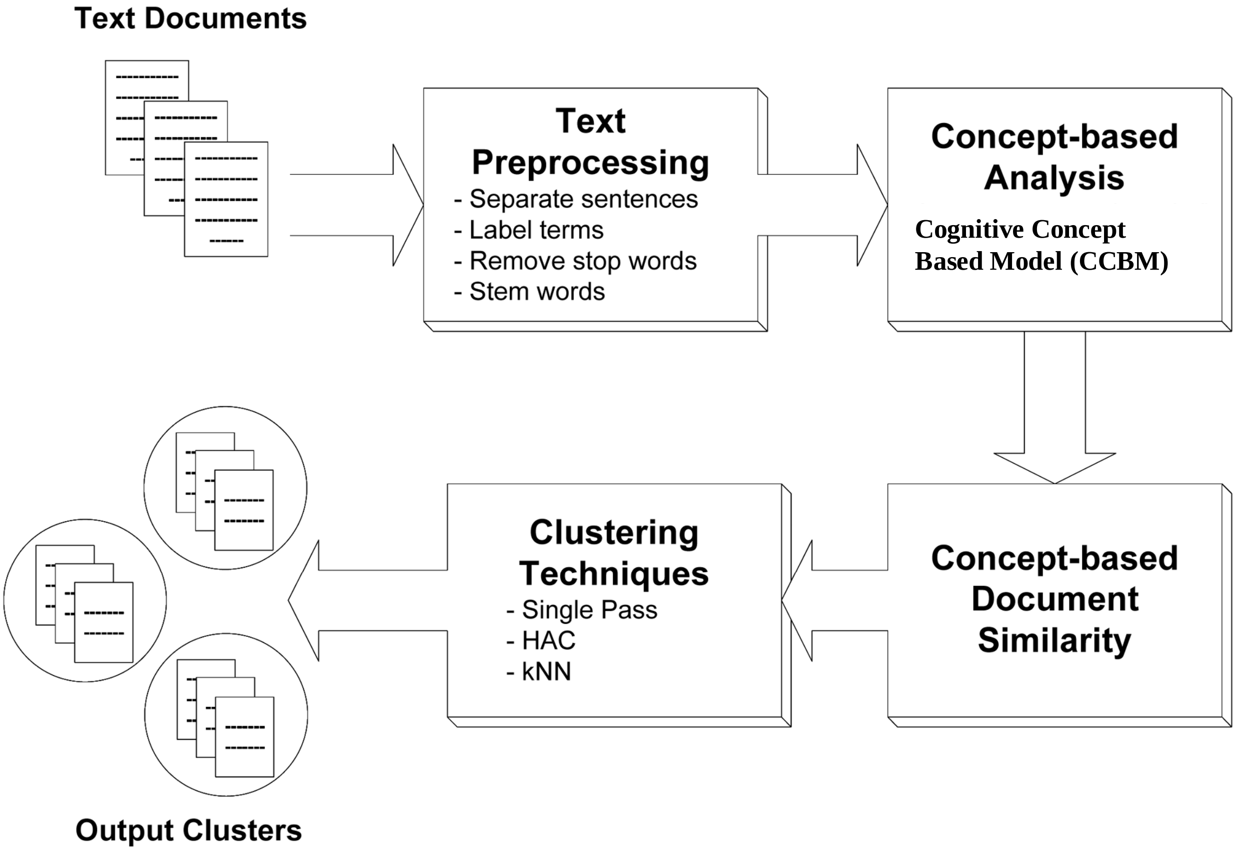


Figure 2.The proposed concept based mining model

Text classification system usually separated into three main phases as follows.

1. Text preprocessing and feature selection phase that makes the dataset more compatible and applicable to train the text classifier.
2. Text classifier phase that use to classify dataset into different classes.
3. Evaluation phase to show the performance of the used classification algorithm.

In concept based model, the term concept is used to describe a labeled term. A term can be either a phrase or a word. The word term is used to indicate both the verb and the argument. A term can have more than one semantic role in the same sentence. This implies that a single term can be an argument to more than one verb in the same sentence. Such terms are said to play more important roles that contribute to the meaning of the sentence.

We used Neural learning method as a preprocessing and analyzing steps which applied on social media document before doing the clustering and then we compared the proposed CCBM model and other concept based methods. The result shows that the proposed method performed better performance than other features selection methods.

IV. EXPERIMENTAL EVALUATIONS

The proposed approach provides superior clustering techniques by incorporating cognitive concept based model along with neural learning techniques on social media text documents. The cluster purity of each generated cluster is evaluated for the proposed model. The average cluster purity enough for the quality of clustering process.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

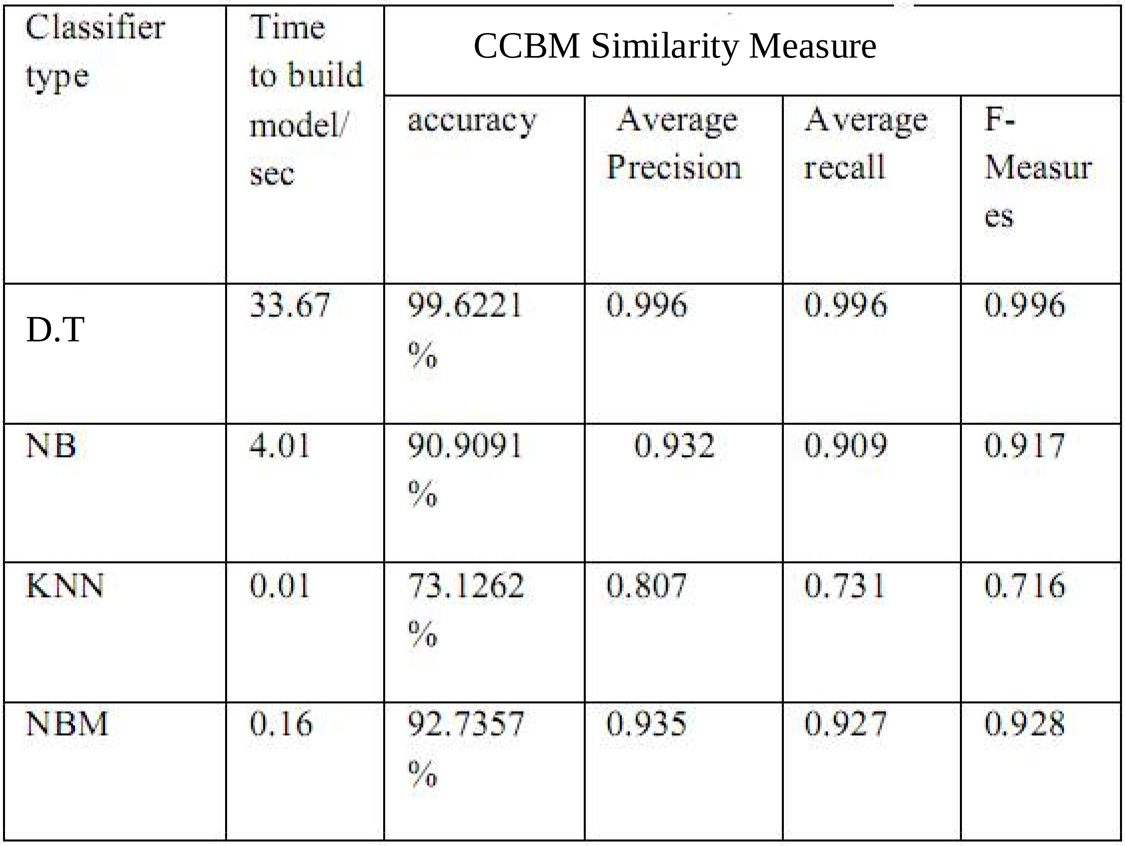
The effectiveness of concept analysis for determining an accurate measure of the similarity between documents, we conducted extensive sets of experiments using the concept-based term analysis and similarity measure. In the social media documents data sets, the text directly is analyzed, rather than, using metadata associated with the text documents. To compare the performance between different classification algorithm (decision tree, K-nearest neighbors(KNN), Naïve Bayesian method and Naive Bayes multinomial classifier) in different situations: using feature selection methods with light stemmer, (khoja stemmer) and using feature selection with full word is adapted at the sentence, document, and corpus levels.

We have evaluated the performance for the classifiers (decision tree, K-nearest neighbors(KNN), Naive Bayesian method and Naïve Bayes multinomial) in terms of precision, recall, accuracy, F-Measures and time to build model as shown in equations 1, 2, and 3. The results are shown in table 1.

$$P_i=TP_i/(TP_i+\ FP_i) \qquad (1)$$

$$R_i=TP_i/(TP_i+\ FN_i) \qquad (2)$$

$$F_i=2P_iR_i/(R_i+P_i) \qquad (3)$$



Classifier type	Time to build model/ sec	CCBM Similarity Measure			
		accuracy	Average Precision	Average recall	F-Measures
D.T	33.67	99.6221 %	0.996	0.996	0.996
NB	4.01	90.9091 %	0.932	0.909	0.917
KNN	0.01	73.1262 %	0.807	0.731	0.716
NBM	0.16	92.7357 %	0.935	0.927	0.928

Table 1. Experiments by taking Concept Similarity

V. CONCLUSION AND FUTURE WORK

The algorithms discussed in the our work propose a novel technique to enhance text clustering in conjunction with concept based mining model which is composed of four components and that improves text clustering quality. The benefit of using concept based model over traditional pure text-based clustering is, the sentence-based concept analysis, document-based concept analysis, the corpus-based analysis, and the concept-based similarity measure create more lucid clusters for improving the clustering efficiency. This will help to visualize the conceptual knowledge with easier interpretation and integration. There are many ways to extend our work with various concepts levels.

REFERENCES

1. R. Feldman and I. Dagan. Knowledge discovery in textual databases (kdt). In Proceedings of First International Conference on Knowledge Discovery and Data Mining, pages 112 - 117, 1995.
2. K.J. Cios, W. Pedrycz, and R.W. Swiniarski, Data Mining Methods for Knowledge Discovery, Kluwer Academic Publishers, 1998.
3. C. C. Aggarwal and C.-X. Zhai, “A survey of text classification algorithms”, in Mining Text Data, New York, USA: Springer, 2012.
4. H. Jin, M.-L. Wong, and K.S. Leung, Scalable Model-Based Clustering for Large Databases Based on Data Summarization, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 11, pp. 1710-1719, Nov. 2005.
5. Guerrero R, Vincent P, Moya A, Victor H, Document Organization using Kohonen's Algorithm. Information Processing and Management, Vol 38, No 1, 2002:79–89.
6. Shan C, Damminda A, et al. (2005) Building an adaptive hierarchy of clusters for text data. International Conference on Computational Intelligence for Modeling, Control and Automation, 2005:7–12.
7. Merkl D (1998), “Text classification with self-organizing maps: Some lessons learned”, Neuro-computing, vol. 21, no. 1–3, 61–77,1998.
8. C. Fillmore, “The Case for Case”, Universals in Linguistic Theory, Holt, Rinehart and Winston, 1968.
9. D. Gildea and D. Jurafsky, “Automatic Labeling of Semantic Roles,” Computational Linguistics, vol. 28, no. 3, pp. 245-288, 2002.
10. G. Salton and M.J. McGill, “Introduction to Modern Information Retrieval”, McGraw-Hill, 1983.
11. G. Salton, A. Wong, and C.S. Yang, “A Vector Space Model for Automatic Indexing”, Comm. ACM, vol. 18, no. 11, pp. 112-117, 1975.
12. P. Kingsbury and M. Palmer, “Propbank: The Next Level of Treebank,” Proc. Workshop Treebanks and Lexical Theories, 2003.
13. Salloum, S. A., Al-Emran, M., &Shaalan, K., “A Survey of Lexical Functional Grammar in the Arabic Context”, Int. J. Com. Net. Tech, 4(3), 2016.
14. Al Emran, M., &Shaalan, K. , “A Survey of Intelligent Language Tutoring Systems”, In Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on (pp. 393-399), IEEE, 2014.
15. Al-Emran, M., Zaza, S., &Shaalan, K., “Parsing modern standard Arabic using Treebank resources”, In Information and Communication Technology Research (ICTRC)(pp. 80-83), IEEE, 2015.
16. SØRENSEN, H. T., Sabroe, S., & OLSEN, J., “A framework for evaluation of secondary data sources for epidemiological research”, International journal of epidemiology, 25(2), 435-442, 1996.
17. Zhang, J. Q., Craciun, G., & Shin, D. (), “When does electronic word-of-mouth matter? A study of consumer product reviews”, Journal of Business Research, 63(12), 1336-1341, 2010.
18. Kamal, S., &Arefin, M. S., “Impact analysis of facebook in family bonding”, Social Network Analysis and Mining, 6(1), 1-14, 2016.
19. Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D.), “Facebook profiles reflect actual personality, not self-idealization”, Psychological science, 2010.
20. Chen, X., Vorvoreanu, M., &Madhavan, K., “Mining social media data for understanding students’ learning experiences”, IEEE Transactions on Learning Technologies, 7(3), 246-259, 2014.
21. Buettner, R., “Predicting user behavior in electronic markets based on personality-mining in large online social networks”, Electronic Markets, 1-19, 2016.
22. Berry Michael, W., “Automatic Discovery of Similar Words- Survey of Text Mining: Clustering, Classification and Retrieval”, Springer Verlag, New York, 200, 24-43, 2004.
23. P. Sollich, A. Krogh, “Learning with ensembles: How over-fitting can be useful”, Advances in Neural Information Processing Systems 8, Denver, CO, MIT Press, Cambridge, MA, 1996, pp. 190–196.