International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Text Categorization and graphical representation using Improved Markov Clustering

Hemanth Somasekar[1]* **Kavya Naveen**[1]

[1]*Department of Computer Science & Engineering, RNS Institute of Technology, India*
* Corresponding author's Email: hemanth.shantha@gmail.com

Abstract: Text categorization means dividing a set of input documents into the two or more classes to which these documents belong. Because of increase in availability of data in digital form in large amount, it becomes necessary to organize it. Feature extraction is the crucial step in text classification. Most of the existing text classifiers are lacking in finding out the relations among the terms. In this research work proposed text clustering of Improved Markov Clustering Model (IMCM) and text classification in which the nonlinear relations among the terms are also considered. This approach is scalable to huge dataset also and its classification power is not affected if relations among terms are large. The proposed IMCM method improved the text document clustering significantly and helped to avoid the human interaction in text clustering. Experimental results have shown that proposed system is 91% accurate for 8 categories and decreases, as increase the classes from 8 to 12, from 96% to 94%, respectively. Also, compared proposed model with existing Naive Bayes and k-Nearest Neighbor classifiers. Experimental results show that the proposed model is more accurate than these two classifiers. The better results demonstrated that presented system is developed effectively.

Keywords: Domain ontology graph, Improved Markov clustering model, K-nearest neighbor, Naïve bayes, Text mining, Text classification.

1. Introduction

In recent years, computerized organization of electronic files is treated as a major study in the domain of computer science [1]. Text files have turn out to be one of the most common vicinity types of statistical repository, mainly due to the elevated reputation of the World Wide Web (WWW) [2]. The main sources of text data generations include websites, newsgroup discussions, forum messages and emails. The amount of digital information is created and used is steadily growing along with the development of sophisticated hardware and software. This has increased the need for powerful algorithms that can interpret and extract interesting knowledge from these data. Data mining is a technique that has been successfully exploited for this purpose. Text mining, a category of data mining, considers only digital documents or text. Text clustering is the process of grouping text or documents such that the document in the same cluster are similar and are dissimilar from the one in other clusters [3, 4]. Majority of the text clustering paradigm employs bag-of-words approach, where each distinct term present in the documents collection is considered as a feature for the document's representation [5]. A data clustering is the divides the documents into two groups such as the closeness among the documents of the same gathering is augmented and the comparability among the data of various groups is minimized [6]. Document clustering idea is utilized as a part of different regions like data recovery, content mining and so on.

Text document classification is the active research area of text mining in which assigning of text documents into classes or categories is done [7]. These text documents include letters, newspapers, articles, blogs, proceedings, journal papers, etc. Text categorization means dividing a set of input documents into the two or more classes to which these documents belong. Because of increase in availability of data in digital form in large amount, it becomes necessary to organize it [8]. Text classification techniques can be divided into two categories: supervised document classification and unsupervised document classification (or document clustering). Supervised classification is one in which for defining the classes and classifying the documents, an external mechanism (e.g., human feedback) provides the information. Supervised machine learning techniques like Support Vector Machine, k-Nearest Neighbors, Naive Bayes, and Decision Tree are applied frequently in text classification [9]. In exiting research works consists of several issues such as in data clustering complexities in high dimensionality of the dataset, and difficult to understand the ability of the cluster description. Also, the result of clustering depends strongly on a set of parameters. Moreover, in most cases, user cannot know in advance how these parameters will affect the final result of the algorithm application. This reduces the computerization potential of such solutions, and also requires that user thoroughly understand both the problem being solved and the mathematical features of the algorithm [10]. To overcome these limitations, the present research work proposes the Improved Markov Clustering Model. Text classification process is divided into two phases: training phase and testing phase. The proposed Improved Markov Clustering Model (IMCM) is cluster the text documents after the graph based text representation. With the help of this proposed method avoids the human interactions and improves the text clustering efficiency significantly.

The rest of the paper is organized as follows: In Section 2 introduces the literature survey of the graph based text clustering and classification. In Section 3 describes the graph based text clustering of proposed research work using improved markov model. The experimental results will be listed and discussed in Section 4. Finally, conclude the paper in Section 5.

2. Literature review

A. Skabar, and K. Abdalgader [11] presented new fuzzy clustering techniquee used for text clustering. This method deals the relational input data and input data were form of a square matrix of pairwise similarities between data objects. The fuzzy technique employs the graph representation of the data. This algorithm recognizes the overlapped clusters significantly but, the proposed technique provided poor results in different bench mark datasets.

A.K. Sangaiah, A.E. Fakhry, M. Abdel-Basset, and I. El-henawy, [12] presented unsupervised, semi-supervised as well as semi supervised with dimensionality reduction clustering methods were used for Arabic text documents clustering and classification. In preprocessing step, all stop words were removed and obtain the root term in each documents. The weighting method provided the weighted values of each and every terms in the documents. Here, the dimensionality reduction technique improved the clustering accuracy. However, increasing the ratio of reduction can sometimes destroy the important terms.

S. Karol, and V. Mangat [13] proposed FCM and K-Means with PSO algorithm that was used for clustering of text documents. Text Document Clustering refers to the clustering of related text documents into groups based upon their content. It is a fundamental operation used in unsupervised document organization, text data mining, automatic topic extraction, and information retrieval. Fast and high-quality document clustering algorithms play an important role in effectively navigating, summarizing, and organizing information. With the help of proposed method accurately clustered the high dimensional data and decreased the overlapping data. But, increased the number of iterations search for the global optimum solution.

H. Zhao, H.S. Salloum, Y. Cai, and J.Z. Huang. [14] presented a novel integrated technique for ensemble subspace clustering of high dimensional sparse text data. The proposed method employs two-level feature representation of text data to generate clusters from subspaces. The proposed clustering methods to increase the robustness of the clusters. This method depends on topic modelling to get the two-level feature representation of text data and to generate different ensemble components. This proposed technique was implemented in real life datasets. While some of these data sets are easy to cluster, others are hard, and some others contain unbalanced data.

J. Jiang, X. Yan, Z. Yu, J. Guo, and W. Tian. [15] proposed a Chinese expert name disambiguation approach based on the semi-supervised graph clustering. The utilization of the attribute correlation, two types of pairwise constraints have been established to act as semi-supervised information, which has been integrated into the initial function that has been built to obtain the final objective function. Finally, with the adoption of the attribute correlation as the semi-

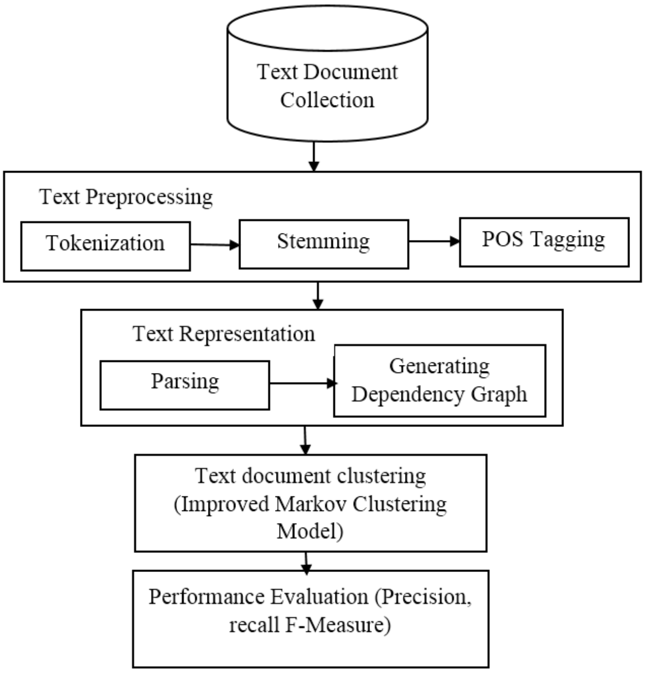
Text Document Collection
Text Preprocessing
Tokenization
Stemming
POS Tagging
Text Representation
Parsing
Generating Dependency Graph
Text document clustering (Improved Markov Clustering Model)
Performance Evaluation (Precision, recall F-Measure)

Figure. 1 Proposed architecture of graph based text representation

supervised constraint, construct an expert disambiguation model by applying the graph-based clustering approach to get the solution of the model through the kernel-based method for the purpose to achieve expert name disambiguation. This technique was applicable for only particular language not a language independent.

3. Proposed research work

Text categorization means dividing a set of input documents into the two or more classes to which these documents belong. Because of increase in availability of data in digital form in large amount, it becomes necessary to organize it. Feature extraction is the crucial step in text classification. Most of the existing text classifiers are lacking in finding out the relations among the terms. The proposed research work of text document clustering of IMCM method is improves the text clustering efficiency. The proposed methodology includes five steps, namely, data collection, document preprocessing, text representation, applying text clustering algorithm and finally performance evaluation as shown in Fig. 1.

3.1 Text dataset acquisition

In this section, would test proposed high-order representation structure and HOCOClu co-clustering on real data sets (20Newsgroups and Reuters-21578). Twenty Newsgroups dataset is downloaded from Jason Rennie’s page. This dataset is a collection of 20,000 newspaper documents approximately, partitioned in 20 categories. This dataset is freely available. A filtered the documents of only 12 classes, i.e., Advertisement, Automobile, Computers, Cryptography, Electronics, Games, Medical, Politics, Religion, Science, Graphics, and Windows for research from these 20 categories dataset. Each file contains on an average of 70 words.

3.2 Text preprocessing

Pre-processing consisting of procedure that transforms the text data in the document to a structure template for text mining. The main goal of processing is to get basic features or key terms from online news text documents and enhance the relevancy between words and document and the relevancy between words and category. A document almost contains number of not necessary words that could negatively affect the clusters of the document. Here, the text preprocessing is majorly includes splitting the documents into sentences, stemming, tokenization and tagging.

- Initially, splits the documents into sentences after that applying stemming process, to reduction of inflected words to their stem, root or base form or in general, a written word form. Stemming is used in determining domain vocabulary in domain analysis. The Porter stemming algorithm is utilized to stem the words in original 20 newsgroup documents.

- The tokenization process is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes the input data for further processing such as parsing or text mining. Tokenization is useful in linguistics, where it is a form of text segmentation, and in computer science, where it forms part of lexical analysis.
- Finally, the part of speech tagging is distributed in all parts of the language of a number of passes. Distinguishable words are grammatically unambiguous early by using a list of English words (a lexical search), and the layers in the syntax. Tagging of ambiguous words is deferred to passes dealing with clausal patterns to utilize the context and enable accurate POS tagging.

3.3 Text representation

A graph based text representation objective is improving the text clustering results. The contributions of this research lies in the method of constructing a graph based text representation which includes extracting noun phrases and chooses the most frequent phrases (top sentences) and convert them into meaningful sentences using the graph based text representation. The text representation consists of two steps such as parsing and generating dependency graph.

3.3.1. Parsing

Parsing is the process of resolving (a sentence) into its component parts and describe their syntactic roles. The main objective of the syntactic parsing region is to recognize a sentence and designation a grammatical structure to it, namely the parse tree. The sentences are parsed to identify syntactic structure. Furthermore, the parser must be having ability to deal efficiently with the problem of ambiguity, where one sentence or words that may have more than one parse. Parsing algorithms dependent on a grammar which is declarative formalities and it can be calculated in several possible ways. The goal of a parser is to analyze an input sentence and output the corresponding (most preferred) parse tree. From the experiences of previous studies, the Stanford parser is used to obtain word dependencies.

3.3.2. Generating dependency graph

In this step, representations of each document as a dependency graph were performed where each node corresponds to a word that can be seen as a meta description of the document. And use the edges between nodes to capture the semantic relations between the pair word. Dependency graph is projective because when all the words written in a linear arrangement. Edges can be drawn without crossing over the words. This is equivalent to saying that word and all the grandchildren (dependents and dependents have dependents, etc.) constitutes a continuum of words in a sentence. After parsing convert the result of the parser to the dependency graph.

3.4 Text document clustering

Clustering classifies the documents into several classes based on the topics. Therefore, each class has one topic. Stressed that one of the main tasks in the text mining is text clustering. The primary aspect of algorithms in clustering contains compactness and isolation of clusters. Markov models have been widely utilized for modelling user's text documents behavior. In this work proposed an improved Markov clustering-based method to increase a Markov model's accuracy in representing a collection of text documents.

3.4.1 Improved Markov clustering model

The Markov Clustering Model (MCM) algorithm is short for the Markov Cluster Algorithm, a fast and scalable unsupervised cluster algorithm for graphs (also known as networks) based on simulation of (stochastic) flow in graphs. The major benefit of this clustering algorithm is simple, fast, scalable and adaptable. There are several ways to speed up the markov model clustering. With the help of IMCM method reducing the size of the data structures employed, and hence also reduce memory demands.

Let us consider the distribution of fluxes and the transition quantities of the fluxes are denoted as M and$T$, respectively and $m_{i,j}$ and $t_{i,j}$ represent the flux flowing from v_j to v_i and the quantity of flux flowing from v_i to v_j in an iteration step. The initial M and T are derived from the adjacency matrix A of a network, is shown in Eq. (1),

$$M_{i,j}^{0} = T_{i,j}^{0} = \frac{A_{i,j}}{\sum_k A_{k,j}} \tag{1}$$

Based on the matrices M and T, a typical Markov clustering algorithm is used to introduce the expansion, inflation, and pruning operators of MCL. First, the expansion operator is used to spread fluxes in a network. It is implemented by multiplying the matrix by the canonical transition matrix$T$, as shown in Eq. (2). Then, the inflation operator is used to enlarge the inhomogeneity and prevent M from converging to the principal eigenvector of$T$. Such an operator raises every entry in M to the power of r and then normalizes the columns as shown in Eq. (3). As all of the entries in$M$are less than or equal to 1, the inhomogeneity in each column is enlarged. In other words, this operator aims to strengthen the strong flows and weaken the weak flows. Finally, the pruning operator accelerates the rate of convergence and reduces non-zero entries to save memory by removing entries under a pre-established threshold. Such thresholds are based on the average and maximum values within columns.

$$M^{st+1} = M^{st} \times T \tag{2}$$

$$M^{ts+1} = \frac{\left[M_{i,j}^{ts+1}\right]^r}{\sum_k \left[M_{k,j}^{ts+1}\right]^r} \tag{3}$$

In each iteration of IMCM, the expansion, inflation, and pruning operators are executed alternately. With the iteration going on, most of the vertices will find an attractor, to which all of their fluxes flow. Each column of M has only one positive value when M converges. The row indexes of those positive values are community labels of the corresponding vertices of each column. The vertices, whose fluxes flow to the same attractor, are clustered to a community.

The proposed IMCM clustering method avoids the overlapping text clusters, decreases the more memory usage. It significantly clusters the high dimensional data and avoids the human interaction.

4. Experimental result and discussion

The proposed algorithm for text classification is implemented and compared with Naive Bayes and k-Nearest Neighbor classifier. Naive Bayes and k-Nearest Neighbor classifiers are implemented in Python using the inbuilt library “sklearn.” In k-Nearest Neighbor algorithm, ten nearest neighbors are considered for measuring the distances in classification. The three classifiers are implemented using 8, 10, and 12 classes or categories to measure the performance of classifiers effectively. It is done to evaluate and compare the effect of number of categories on the classification power of the classifier. To evaluate the power of classifiers, the comparison is done using precision, recall, and f-measure. The estimation has been done for these parameters using TP, FP, FN, and TN values, where TP refers to true positive, TN is true negative, FP is false positive and FN is false negative. The calculation of parameters is described below,

Precision: Precision is the proportion of the examples which truly have class x among all those which are classified as class x. This is defined in Eq. (4),

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

Recall: It is a measure of the ability of a prediction model to select instances of a certain class from a data set also called as sensitivity and corresponds to the true positive rate. It is defined in Eq. (5),

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

Accuracy: Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. The accuracy is directly proportional to true results, consider both true positives and true negatives among the total number of cases scrutinized. The parameter of accuracy is calculated in Eq. (6),

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{6}$$

F-Measure: It is the measure of harmonic mean of precision and recall. It gives the closeness between precision and recall. It is defined by as mentioned in Eq. (7),

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

It is the measure of harmonic mean of precision and recall. It gives the closeness between precision and recall. It is defined by as mentioned in Eq. (3) considered the number of categories N = 8 for text classification. These are Automobile, Electronics, Religious, Sports, Medical, Cryptography, Science, and Politics. Then, the performance is evaluated using precision, recall, f-measure and accuracy. Table 1 shows the values of precision, recall, and f-measure for different models using number of classes N = 8.

Table 1. Performance comparison of N=8

Parameters	Method	Automobile	Electronics	Sports	Science	Politics	Computer	Average
Precision	NB	0.83	0.94	0.94	0.85	0.87	0.85	0.84
	KNN	0.61	0.61	0.83	0.71	0.95	0.89	0.89
	DOG	0.84	0.84	0.92	0.84	0.9	0.91	0.93
	Proposed IMCM	0.92	0.9	0.91	0.97	0.93	0.9	0.98
Recall	NB	0.85	0.75	0.75	0.87	0.83	0.94	0.89
	KNN	0.89	0.71	0.83	0.95	0.61	0.83	0.91
	DOG	0.91	0.84	0.84	0.9	0.84	0.92	0.94
	Proposed IMCM	0.92	0.98	0.98	0.97	0.95	0.96	0.98
F-Measure	NB	0.79	0.84	0.75	0.65	0.79	0.85	0.83
	KNN	0.81	0.73	0.89	0.78	0.61	0.71	0.61
	DOG	0.92	0.91	0.81	0.83	0.84	0.84	0.84
	Proposed IMCM	0.98	0.94	0.95	0.96	0.91	0.98	0.94
Accuracy	NB	0.75	0.73	0.75	0.69	0.74	0.83	0.75
	KNN	0.88	0.71	0.71	0.71	0.83	0.61	0.88
	DOG	0.93	0.89	0.84	0.82	0.91	0.84	0.93
	Proposed IMCM	0.9	0.94	0.85	0.98	0.94	0.82	0.96

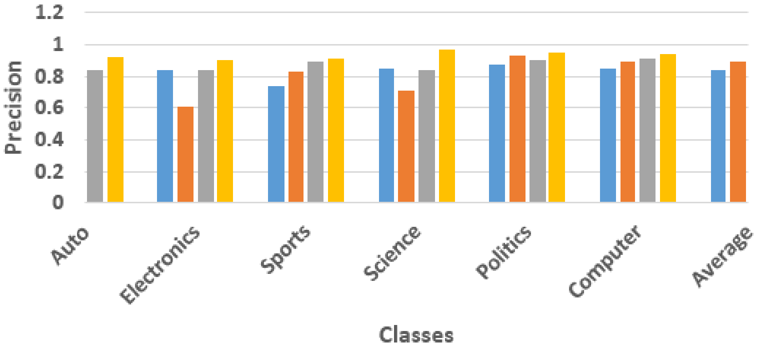

Figure. 2 Precision performance N=8

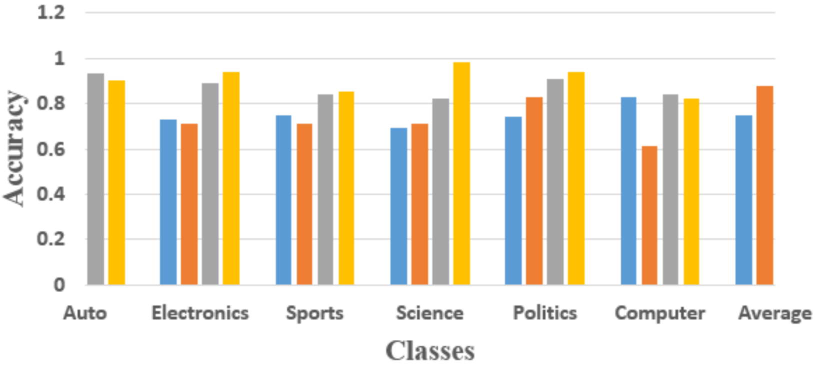

Figure. 3 Accuracy performance of N=8

Fig. 2 gives the representation for comparison of precision for different classifiers for different classes. The accuracy power for DOG is 30%, while those of Naive Bayes are 75%, k-NN is 88% and proposed IMCM method achieves 96% of accuracy. This f-measure value shows that the DOG proposed model performs better than other two classifiers. Here, the DOG method shows the lowest accuracy. Proposed model shows maximum of 98% accurate results for class Sports as well as Science and minimum of 91% for class Politics. This result shows that proposed model gives better result as compared to the other three techniques.

Fig. 2 represents the precision performance of the existing and proposed methods with respect to different classes. Here, the precision of the NB method represents the lowest precision approximately 84%, KNN achieves 89% of precision, DOG shows 93% of precision and proposed IMCM achieves 98% of precision. Compare to the other existing technique the proposed IMCM method shows better results. Fig. 3 represents the accuracy performance of the existing and proposed method with respect to the different classes and class size is N=8. Compare to the exiting NB, KNN, DOG model the proposed IMCM method shows the better results.

Considered the number of categories N = 12 for text classification. These are Automobile, Electronics, Religious, Sports, Medical, Windows. Then, the performance is evaluated using precision, recall, and f-measure. The accuracy power for DOG is 89%, while those of Naive Bayes are 73%, k-NN is 71% and proposed IMCM achieves the 94% of accuracy. The f-measure value shows that the IMCM proposed model performs better than other three classifiers. The k-NN has lowest accuracy and proposed model shows maximum of 94% accurate results. The class science, computer, politics shows the maximum of 98% accuracy. The class electronics shows the minimum of 94% of accuracy. This result shows that proposed model gives better result as compared to the other three techniques. This comparison can also be expressed using graphical representation.

Fig. 4 shows the graphical representation for comparison of the three techniques such as NB, KNN, DOG, and proposed IMCM method for text classification. The graphical representation of average f-measure for all the three classifiers. These show that with increase in number of categories, the accuracy of the classifier decreases. Also, observed that the f-measure value decreases for every classifier with increase in value of N. The F-measure of the NB and KNN method shows the minimum results and proposed IMCM indicates the maximum results than the other existing methods.

Table 2. Performance comparison of N=12

Parameters	Method	Automobile	Electronics	Sports	Science	Politics	Computer	Average
Precision	NB	0.64	0.65	0.77	0.65	0.63	0.84	0.74
	KNN	0.79	0.79	0.68	0.71	0.71	0.73	0.61
	DOG	0.93	0.91	0.8	0.84	0.84	0.92	0.84
	Proposed IMCM	0.98	0.9	0.95	0.97	0.95	0.96	0.98
Recall	NB	0.67	0.69	0.65	0.75	0.65	0.64	0.79
	KNN	0.75	0.71	0.73	0.65	0.79	0.73	0.61
	DOG	0.8	0.84	0.84	0.84	0.85	0.82	0.82
	Proposed IMCM	0.97	0.98	0.98	0.98	0.92	0.96	0.98
F-Measure	NB	0.63	0.69	0.62	0.85	0.79	0.75	0.64
	KNN	0.77	0.71	0.78	0.68	0.61	0.869	0.73
	DOG	0.84	0.84	0.83	0.73	0.84	0.81	0.81
	Proposed IMCM	0.82	0.9	0.96	0.8	0.91	0.82	0.94
Accuracy	NB	0.75	0.72	0.74	0.69	0.85	0.75	0.73
	KNN	0.68	0.75	0.85	0.71	0.71	0.71	0.71
	DOG	0.83	0.88	0.82	0.92	0.84	0.84	0.89
	Proposed IMCM	0.9	0.94	0.85	0.91	0.98	0.92	0.94

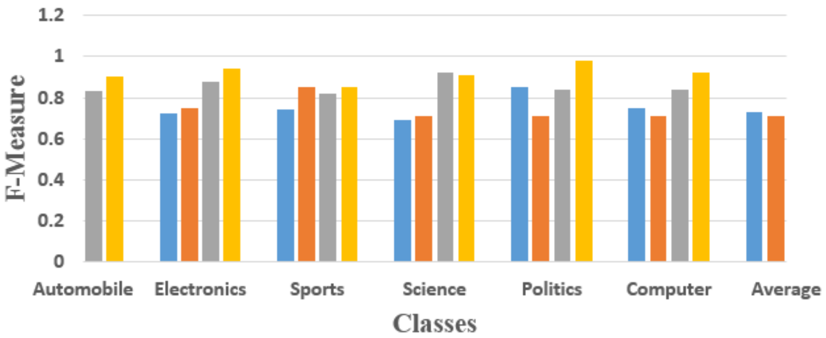

Figure. 4 F-measure performance of N=12

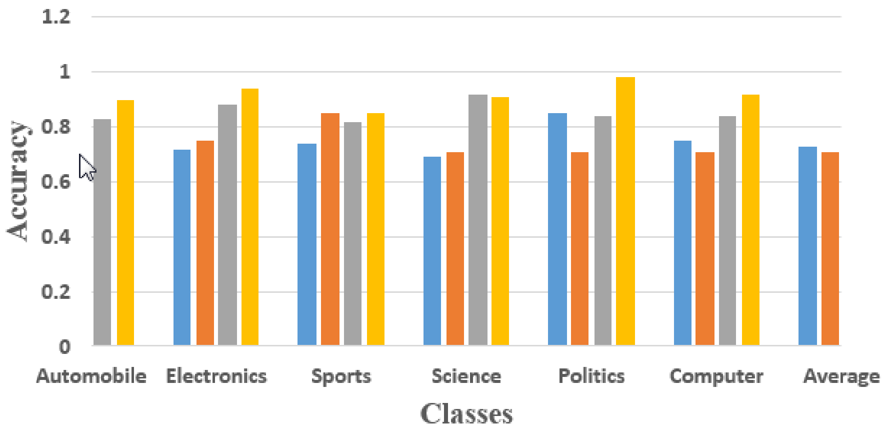

Figure. 5 Accuracy Performance of N=12

Table 3. Comparison of different Text document clustering and classification methods

Existing Work	Method	Parameters
J.N. Liu, Y.L. He, E.H. Lim, and X.Z. Wang [16]	DOG model for Chinese text classification	Precision: 92.5%, Recall: 91.0%, F-Measure: 92.3%
Ramkumar, A. Sudha, and B. Poorna. [17]	Text document clustering using dimension reduction technique	Recall :90%, Precision : 96%, Accuracy : 91.9%
Kang, Jiayin, and Wenjuan Zhang. [18]	Hybrid FCM and PSO algorithm for text document clustering	F-Measure: 86.01
P. Subba, S. Ghosh, and R. Roy [19]	Single document extractive text summarization using FCM and K means method.	K-mean recall : 79.2%, precision: 91.4%, F-score : 71.7 FCM recall: 59.5, precision: 68.3%, F-score: 61.5.
M.B. Revanasiddappa, B.S. Harish, and S.V.A. Kumar, [20]	Kernel Probabilistic C means using text document clustering	Accuracy : 91.46%
Proposed work	IMCM model using document clustering	Precision:98% Recall:98% Accuracy:94% F-Measure:94%

Fig. 5 represents the accuracy of the existing methods and proposed IMCM method with respect to the different classes and class size is N=12. The NB and KNN method shows minimum accuracy in all the classes. The proposed IMCM represents the high accuracy and better efficiency in all the classes. If the number of classes is less than the accuracy is high and if the number of classes are increased, then the accuracy is bit decreased. But the proposed IMCM method indicates the better results.

Table 3 represents the text document clustering techniques of existing and proposed methods. The Liu, J.N. Liu, Y.L. He, E.H. Lim, and X.Z. Wang [16] proposed Chinese text classification using DOG model. This work focuses only Chinese text data no other kinds of data then only achieve 92.5% of precision, 91.0% of accuracy, and F-Measure 92.3% but, proposed algorithms focuses all kinds of text data. Ramkumar, A. Sudha, and B. Poorna. [17] proposed dimension reduction technique based text document clustering. In this paper experimental analysis taken BCC sports dataset therefore it concentrates sports categories only. But proposed graph based IMCM clustering method is uses all different kinds of categories of text data such as sports, computer, politics, science, automobile, electronics, etc. Kang, Jiayin, and Wenjuan Zhang [18] proposed FCM and PSO based text clustering. In this paper the partition is done such that patterns within a group are more similar to each other than patterns belonging to different groups. Thus, difficult to identify the text belongs which cluster therefore the performance is decreased. P. Subba, S. Ghosh, and R. Roy [19] proposed FCM and K means methods used for text summarization. In this work, single document is extracted and text summarization is done not for multiple documents. M.B. Revanasiddappa, B.S. Harish, and S.V.A. Kumar [20] proposed KPCM method text clustering. Here, high dimensional features are employed for similar text clustering so, computational time is high. However, the proposed IMCM model is overcome the existing problems and achieves the 98% of precision and recall respectively, 94% of accuracy and F-Measure shows the 94%.

5. Conclusion

Text mining, also known as text data mining or knowledge discovery process from the textual databases, generally, is the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents. All the extracted information is linked together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. This paper presents a new framework design for document clustering using graph based Improved Markov Clustering Model. Initially review the several existing work and present a method of document clustering using frequent term set. The proposed IMCM based clustering method shows better results than the previous text document clustering methods. Here, IMCM clustering methods avoids the overlapped clusters, decreases the CPU time and less memory consumption. An experimental analysis shows that the texts are taken from 20Newsgroups and Reuters-21578. The results indicated as the accuracy of the proposed method is 94%, recall is 98%, precision is 98%, and the F-Measure achieves 94% in all the different text classes. So, the proposed method experimental performance is better than the existing techniques. In future, hybrid clustering method will use for multiple language based text document clustering.

References

[1] C. Qimin, G. Qiao, W. Yongliang, and W. Xianghua, "Text clustering using VSM with feature clusters", *Neural Computing and Applications*, Vol.26, No.4, pp.995-1003, 2015.

[2] S.N.B. Bhushan, and A. Danti, "Classification of text documents based on score level fusion approach", *Pattern Recognition Letters*, Vol.94, pp.118-126, 2017.

[3] S.C. Punitha and M. Punithavalli, "Performance evaluation of semantic based and ontology based text document clustering techniques", *Procedia Engineering*, Vol.30, pp.100-106, 2012.

[4] E. Rahimtoroghi and A. Shakery, "Wikipedia-based smoothing for enhancing text clustering", In: *Proc. of International Conf. on Information Retrieval*, pp.327-339, 2011.

[5] K.K. Bharti and P.K. Singh, "Opposition chaotic fitness mutation based adaptive inertia weight BPSO for feature selection in text clustering", *Applied Soft Computing*, Vol.43, pp.20-34, 2016.

[6] K. Shi and L. Li, "High performance genetic algorithm based text clustering using parts of speech and outlier elimination", *Applied Intelligence*, Vol.38, No.4, pp.511-519, 2013.

[7] E.V. Kotelnikov and M.V. Pletneva, "Text sentiment classification based on a genetic algorithm and word and document co-clustering", *Journal of Computer and Systems Sciences International*, Vol.55, No.1, pp.106-114. 2016.

[8] Z. Wang and Z. Liu, "Graph-based KNN text classification", In: *Proc. of 7th International Conf. on* Fuzzy Systems and Knowledge Discovery, Vol.5, pp.2363-2366, 2010.

[9] L.H. Lee, C.H. Wan, R. Rajkumar, and D. Isa, "An enhanced Support Vector Machine classification framework by using Euclidean distance function for text document categorization", *Applied Intelligence*, Vol.37, No.1, pp.80-99, 2012.

[10] E.K. Mikhina and V.I. Trifalenkov, "Text clustering as graph community detection", *Procedia Computer Science*, Vol.123, pp.271-277, 2018.

[11] A. Skabar and K. Abdalgader, "Clustering sentence-level text using a novel fuzzy relational clustering algorithm", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 25, No.1, pp.62-75, 2013.

[12] A.K. Sangaiah, A.E. Fakhry, M. Abdel-Basset, and I. El-henawy, "Arabic text clustering using improved clustering algorithms with dimensionality reduction", *Cluster Computing*, pp.1-15, 2018.

[13] S. Karol and V. Mangat, "Evaluation of text document clustering approach based on particle swarm optimization", *Open Computer Science*, Vol.3, No.2, pp.69-90, 2013.

[14] H. Zhao, S. Salloum, Y. Cai, and J.Z. Huang, "Ensemble subspace clustering of text data using two-level features", *International Journal of Machine Learning and Cybernetics*, Vol.8, No.6, pp.1751-1766, 2017.

[15] J. Jiang, X. Yan, Z. Yu, J. Guo, and W. Tian, "A Chinese expert disambiguation method based on semi-supervised graph clustering", *International Journal of Machine Learning and Cybernetics*, Vol.6, No.2, pp.197-204, 2015.

[16] J.N. Liu, Y.L. He, E.H. Lim, and X.Z. Wang, "Domain ontology graph model and its application in Chinese text classification", *Neural Computing and Applications*, Vol.24, No.3-4, pp.779-798, 2014.

[17] A.S. Ramkumar and B. Poorna, "Text document clustering using dimension reduction technique", *International Journal of Applied*

Engineering Research, Vol.11, No.7, pp.4770-4774, 2016.

[18] J. Kang and W. Zhang, “Combination of fuzzy C-means and particle swarm optimization for text document clustering”, In: *Proc. of International Conf. On Advances in Electrical Engineering and Automation*, pp.247-252, 2012.

[19] P. Subba, S. Ghosh, and R. Roy, “Partitioned-Based Clustering Approaches for Single Document Extractive Text Summarization”, In: *Proc. of International Conf. on Mining Intelligence and Knowledge Exploration*, pp. 297-307, 2017.

[20] M.B. Revanasiddappa, B.S. Harish, and S.V.A. Kumar, “Clustering Text Documents Using Kernel Possibilistic C-Means”, In: *Proc. of International Conf. on Cognition and Recognition,* pp.127-134, 2018.