Volume 1, Issue 1 (2018)
Article No. 7
PP 1-7

VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL
VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

Survey on Efficient Techniques of Text Mining

Sunita Naik[1], Samiksha Gharat [1], Saraswati shenoy[1], Rohini Kamble[1]
[1](*Computer, VIVA Institute of Technology/ Mumbai University, India,)*

Abstract: *In the current era, with the advancement of technology, more and more data is available in digital form. Among which, most of the data (approx. 85%) is in unstructured textual form. So it has become essential to develop better techniques and algorithms to extract useful and interesting information from this large amount of textual data. Text mining is process of extracting useful data from unstructured text. The algorithm used for text mining has advantages and disadvantages. Moreover the issues in the field of text mining that affect the accuracy and relevance of the results are identified.*

Keywords *–MWO, Consensus, PSO, Text mining, Bisecting K-means*

1. INTRODUCTION

Data mining is the process of sorting through large data set to identify patterns and establish relationship to solve problems through data analysis. The size of data is increasing at exponential rates day by day. Almost every type of organization stored their data electronically. Text mining plays important role in search engine, every text is digitally stored. (Stored in binary form that is 0, 1)Data mining is the mining of the predictive information from database and it is new technology to help companies focus on the very important information in their data bases. It is used to examine the old data to find the information. Since clustering is used and it is one of the popular technique of data mining. It is a task of dividing a data into the number of similar clusters. Means it is task of grouping a set of object in a same group that are similar to each other in the other group. Data clustering technology is to finding the similar hidden pattern from the given data set. It is the method to obtaining the cluster of the item without the class label related to the approximation of the item in one cluster. Clustering is the very big amount of the data set that contains the large number of records with high dimensions. And now a days it used for the identifying useful information from the historical data. The optimization is used to find the global optimization solution. Now a days in real word the optimization problem are dynamic. It will not find the global optimal solution but also find the trajectory of changing optimal solution over dynamic nature.The optimization technique will give the optimal or good solution from the complex optimization problem.

2. Data Mining Techniques

2.1 A Review on Clustering Analysis based on Optimization Algorithm for Data mining [5]

Clustering analysis is one of the important concepts of data mining. It will divide the data into certain classes according to the main attribute of the data set. It has drawback like optimal path, initialization of cluster center. In this after applying k-mean, Bisecting k-mean is applied on obtained cluster. It will find the k number of cluster of the apply data set. Then applying the optimization algorithm it will find the optimize path of the clustering and increase the accuracy of the integrated hybrid algorithm.

In this Bisecting K-mean Technique is used along with PSO and they are good at maintaining final cluster.

2.2 Bisecting K-means Algorithm for Text Clustering [14]

Three steps are used in this, the first one is Pre-processing text, it is easy to compare to natural language documents. The second step is application of text mining Technique, in this the algorithm such as clustering, classification, summarization, information extraction are used. The third step is analysis of text, in this the outputs are analyzed for discovering the knowledge.
This paper gives the idea about basics of text mining.

2.3 Algorithm of Group Members' consensus orienting to Discussion Dynamic Process [6]

To solve this dynamic expansion process, they had proposed a new algorithm of group members' consensus orienting to discussion dynamic process. According to the extraction and clustering of expert's discussion information, experts weight changes dynamically under discussion dynamic process. At the same time the consensus state of group discussion change dynamically. If we claim C1 then, if focus=4, value is 0.1538 and exact consensus vale is 3.3846.

This paper has an algorithm for calculating consensus value based on cluster analysis and the value of modality and the method is feasible and effective.

2.4 Stability of Distributed Adaptive Algorithms I: Consensus Algorithms [7]

Performance analysis (convergence and mean squared error measures) has been pursued under two regimes i.e. fixed gain (aka short memory) or vanishing gain (aka long memory). In vanished gain there are many types of similarities where as in fixed gain there are less similarities. It has two types of noise. The first one is white noise which has equal intensity at different frequencies and second one is colored noise which generates random data.

Since this algorithm is good at removing noise sensation or error outputs so it can be used after applying k-means.

2.5A Modified Particle Swarm Optimization with Dynamic Particles Re-initialization Period [8]

The particle swarm optimization (PSO) is an algorithm that attempts to search for better solution in the solution space by attracting particles to converge toward a particle with the best fitness. In order to overcome problems they have propose an improved PSO algorithm that can re-initialize particles dynamically when swarm traps in local optimum. Moreover, the particle re-initialization period can be adjusted to solve the problem appropriately. The proposed technique is tested on benchmark functions and gives more satisfied search results in comparison with PSOs for the benchmark functions[9]. The PSO has many advantages such as rapid convergence, simplicity, and little parameters to be adjusted. Its main disadvantage is trapping in local optimum and premature convergence. Since the improved PSO technique is good at initializing cluster centre.

2.6 Mussels Wandering Optimization: An Ecologically Inspired Algorithm for Global Optimization [1]

Over the past few year there are various complex optimization problem. To overcome problems of text mining we use Mussels wandering optimization and also compare it with various algorithm to observe which algorithm give better solution. Novel meta heuristic algorithm which is also called as mussels wandering optimization technique is used in this paper.it is inspired by mussels locomotion behavior when they form bed pattern in their habitant.it give more important to the mussels and find their density in habitant.one of the most significant merits of MWO is it provide open frame work to tackle hard optimization problem.

2.7 A Data Clustering Algorithm Based on Mussels Wandering Optimization [2]

The clustering algorithm like k-means algorithm is used to form a cluster. but it have some drawback in searching optimal solution, considering this drawback and limitation. To overcome these drawback in this paper they proposed new algorithm based on k-mean and mussels wandering optimization.
The aim of this algorithm is to reach an optimal solution by mathematically modeling mussels.

Volume 1, Issue 1 (2018)
Article No. 7
PP 1-7

In k-MWO ,each mussels represent a set of center of 'K' classes. the algorithm first initialize 'N' mussels and evaluate each mussels fitness by using squared sum error. according to the fitness value we find the top mussels and update their position in database.

This paper has given the idea of merging MWO with various algorithm and they get accuracy in point tabular form and by combining these two algorithm we can make a full use of global optimization ability of MWO and local search ability of the k-means algorithm.

2.8 A Survey Paper for Finding Frequent Pattern In Text Mining [3]

Text mining is very important method for finding important information from large amount of data.in data mining there three important rule for finding frequent data pattern. First one is frequent pattern and second is association rules .in this paper they used frequent pattern rule for temporal text mining. This technique involve data mining and extracting information. Disadvantage of pattern based method is low frequency and misinterpretation.in this may noisy parameter is discovered, to solve this problem they used term based method.

2.9 Text Mining: Techniques, application and issues [9]

This paper describes review of text mining. Over 80% information is made of unstructured and semi structured information. Content mining is procedure of removing data from huge dataset. By choosing the great strategy we can enhance the speed and lessening the time and efforts which are required to extract the information or content. Some techniques used for text mining are Information Extraction, Information Retrieval, clustering, text summarization. Application of Text Mining are Academic and research field Digital library, Business Intelligence & Social Media.

This paper highlighted the techniques application and issues of text mining. Nowadays application of text mining used in every field. NLP and entity recognition techniques reduced the issues that occur during text mining process. Text mining tools also used in life science i.e. in biomedical field which provides an opportunity to extract important information, their association and relationship among various diseases, species, and genes etc.

2.10 A comparative Analysis of particle swarm optimization and k-mean Algorithm for Text clustering using Nepali word net[10]

This paper discussed about particle swarm optimization and k-means algorithm. Paper portrays investigation of three calculation i.e. k-means, particle swarm optimization and hybrid PSO+ k-means clustering. Clustering is characterized as collection of information into bunches or groups with the goal that the information or record in each group are similar to each group and dissimilar other group. Hybrid PSO +k-means algorithm combines two modules PSO module & k-means module. This will first (hybrid) execute PSO clustering algorithm by global search. PSO will terminate when no of iteration is done. The hybrid PSO algorithm combines the both advantage i.e. globalize searching of PSO and fast coverage of k-means.

This paper Highlighted k-means, bisecting k-means and hybrid PSO+ k-means algorithm. The K-means algorithm was compared with PSO and hybrid PSO+K-means algorithms. Hybrid PSO+K-means performs better than PSO and K-means algorithms. Similarity between two documents need to be computed in a clustering analysis. There are similarity measures are available to compute the similarity between two documents like Euclidean distance, Manhattan distance, cosine similarity etc. among that cosine similarity measurement has been used.

2.11 Review on clustering web data using PSO[11]

This paper described about the clustering technique for web data mining text extraction and clustering are the main challenging tasks .The literature overview a developmental bio inspired swarm intelligence algorithm called as particle swarm optimization for improve result. this algorithm will enhance the efficiency information is conflicting, unstructured and fragmented such issue can be solved by utilizing prepossessing which will raw information into extremely proper arrangement. Subsequent to proposing will apply PSO algorithm on web information for clustering purpose of content utilized for the web text clustering.

This paper highlighted the particle swarm optimization algorithm as well as clustering techniques such as Partition Clustering, Hierarchical Clustering, Density-based Clustering, Grid-based Clustering, model-based

Clustering, and Fuzzy Clustering. Also PSO compared with two other algorithms genetic algorithm and ACO algorithm but PSO gives better result in terms of time, speed and it has low memory requirement & low computational cost.

2.12 A limited Iteration Bisecting k-means for fast clustering large datasets[12]

This paper describes about the bisecting k –means algorithm with compared to k-mean algorithm. About limit no. of iterations. It maintains the clustering quality with limited iteration. They have introduced bisecting k-means which will divide two clusters using k-means with k=2 resulting in two clusters. This bisecting process will continue until getting total no of cluster reaches to k. bisecting k-means is an improvement of k-means in clustering quality as well in efficiency in large dataset. Each two means start with different pair with initial center.

This paper highlighted the limited iteration bisecting k-means for clustering the large dataset. The original version bisecting k means performs multiple runs of two means. The bisecting k-means produces more better and efficient clustering than the k-means.

3. ANALYSIS TABLE

Table 1: Analysis Table

Sr. No.	Title	Technique/Methods	Parameter	Accuracy
1	Mussels Wandering Optimization: An Ecologically Inspired Algorithm For Global Optimization.	Mussels Wondering Optimization.(MWO)	Function ‘f’	Function(f1) : μ(d=20) If μ = 1.5 then the results: Best = 273.99 Mean=1.47e+4
2	A Data Cluster Algorithm Based On Mussels Wandering Optimization.	K-MEAN and Mussels Wondering Optimization(MWO).	DI :- it measure the ratio between distance and diameter of cluster.	DI : Max -0.1128 , Min -0.1009, Mean-0.1021. DBI : Max -0.4375, Min -0.3916, Mean -0.4231.
3	Survey Paper For Finding Frequent Pattern In Text Mining.	frequent pattern rules , extracting information rules.		

4	Mussels wondering algorithm based training of artificial neural network for pattern classification.	In this paper they applied MWO on artificial neural network.	Classification accuracy training time	Classification accuracy : 78.3 training time : 1.48 sec.
5	A Review on Clustering Analysis based on Optimization Algorithm for data mining	Bisecting k-mean and Particle Swarm Optimization (Used to overcome the dependency of method to initialize the cluster).	For calculating Distance between cluster 1 and cluster2 If dist1>dist2 then divide cluster1 into two more cluster, if dist2>dist1 then again divide cluster into two morw cluster	
6	Bisecting k-means Algorithm for Text Clustering	Bisecting k-mean with Time Complexity	To compute two clusters with k=2 and the run time complexity of the algorithm will be O((K-1)IN).	
7	Algorithm of Group Members' consensus orienting to Discussion Dynamic Process	Consensus Building Algorithm	Consensus value of claim CJ is Consensus(c)?LA; x vij, A.i is expert i's weight and Vij is expert i's modality to claim cj .	If we claim C1 then ,if focus=4, value is 0.1538 and exact consensus vale is 3.3846
8	Stability of Distributed Adaptive Algorithms I: Consensus Algorithms	Analysis of a consensus based distributed LMS algorithm under some colored noise assumptions.	If $\mu[\lambda max(L) + max\ k\ \lambda max(Rx,k)] < 2$ This means that for each node E(˜ wk,t) → w∗.	
9	A Modified Particle Swarm Optimization with Dynamic Particles Re-initialization Period	Particle Swarm Optimization		acceleration constants of 1η and 2η is 1.496180 and inertia weight ω = 0.729844 population is 20.maximum iteration is 5000

10	Text mining: techniques application and issues	They used extraction, information retrieval, clustering and text summarization.		
11	comparative analysis of particle swarm and k means algorithm for text clustering using nepali wordnet	They used k means, pso & hybrid pso+ k means algorithm.		For 50 document hybrid pso+ k means gives 6.964 for intra cluster & 0.952 for inter cluster.
12	Review on clustering web data using particle swarm optimization	They use three algorithm PSO,GA, AGO.	Better cost, memory requirement, simplicity etc.	
13	A limited iteration bisecting k means for fast clustering datasets.	They used bisecting k means also describes limited iteration bisecting k means algorithm (LIBKM).	Bisecting k means is better than k means. This will keep the limit of iteration no.	LIBKM will divide 2 clusters using k means with k=2. this will accurate the clustering quality by removing error & validating the cluster.
14	A survey on particle swarm optimization algorithm application in text mining	PSO based data clustering method.	They have compared PSO with GA, SA but PSO gives better result in terms of accuracy & efficiency.	

4. CONCLUSION

This paper presents the significance of text mining and study of techniques used for text mining. Organized Structure with arrangement and clustering techniques are also presented in the survey. The survey paper also include the information of the different data mining algorithm which will give the detailed information about the text mining and it's also clarify the advantages and disadvantages of the data mining. The application of different text mining techniques for unstructured informational collections are reside in the form of text documents. The kind of techniques are permits making a best web engine utilizing database learning to work with filter, wrapper or even ontology. It also described open areas and testing issues explore directions in text mining.

Volume 1, Issue 1 (2018)
Article No. 7
PP 1-7

REFERENCES

[1] Jing An, Qi Kang, Lei Wang, Qidi Wu "Mussels Wandering Optimization: An Ecologically Inspired Algorithm for Global Optimization*" IEEE International Conference on Networking, Sensing and Control.*

[2] Peng Yan, ShiYao Lui, Bing zyao Huang "A Data Clustering Algorithm Based on Mussels Wandering Optimization" *IEEE International Conference 2014.*

[3] Ms.Sonam Tripathi, Asst prof.Tripathi Sharma."A Survey Paper for Finding Frequent Pattern In Text Mining" *International Journal of Advanced Research in Computer Engineering &Technology*(IJRCET)

[4] Ahmed A. Abusnaina, Rosni Abdullah. "Mussels Wandering Optimization Algorithm Based Trainning of Artifical Neural Networks For Pattern Classification" *International Conference on Computing and Information.(ICOCI)2013*

[5] Rashmi P. Dagde, Snehlata Dongre "A Review on Clustering Analysis based on Optimization Algorithm for Data mining". *IJCSN International Journal of Computer Science and Network, Volume 6, Issue 1, February 2017.*

[6] Zhang Zhen, Chen Chao, Chen jun-liang "Algorithm of Group Members' consensus orienting to Discussion Dynamic Process". *IEEE Transaction.*

[7] Victor Solo "Stability of Distributed Adaptive Algorithms I: Consensus Algorithms" *IEEE Transaction 2015*.

[8] Chiabwoot Ratanavilisagul and Boontee Kruatrachue "A Modified Particle Swarm Optimization with Dynamic Particles Re-initialization Period". *Springer International Publishing Switzerland 2014.*

[9] Ramzan Talib, Muhammad kashif Mani, Shaeela Ayesha, Fakeeha Fatima, "Text Mining: Techniques, application and issues", *IJACSA(2016)*

[10] Sarkar, Arindam Roy & B.S Purkayastha," A comparative Analysis of particle swarm optimization and k-mean Algorithm for Text clustering using Nepali wordnet", *IJNLC(June 2014)*

[11] Jayshree Ghorpade-Aher, Roshan Bagdiya,"Review on clustering web data using pso", *International Journal of computer application(December 2014)*

[12] Yu Zhuang, YuMau, Xinchen, "A limited Iteration Bisecting k-means for fast clustering large datasets", *IEEE trust com(2016)*

[13] Rekha Dahiya, Anshima Singh, "A survey on application of particle swarm optimization in Text Mining", *International Journal of Innovative research & development(May 2014)*

[14] Nikita P. Katariya, Prof. M. S. Chaudhari "Bisecting K-means Algorithm for Text Clustering". *IJARCSSE February 2015.*