INTERNATIONAL JOURNAL OF COMPUTER SCIENCE ENGINEERING RESEARCH AND DEVELOPMENT (IJCSERD)

ISSN 2248 – 9363(Print)
ISSN 2248 – 9371(Online),
Volume 3, Number 2, April- May (2013), pp: 46-59

IJCSERD

© PRJ PUBLICATION

AN OPTIMAL UNSUPERVISED TEXT DATA SEGMENTATION USING GENETIC ALGORITHM

N. Silpa[1], Principal Investigator
Prof. V. V. R. Maheswara Rao[2], Scientist Mentor

[1, 2]Shri Vishnu Engineering College for Women, Bhimavaram, AP, India

ABSTRACT

The popularity of information available in electronic forms has been rapidly growing in the last decade, and turn into a golden mount containing extremely unstructured data for the researchers. Extracting interesting information and knowledge from such data creates promising future path into the era of text mining. The roots of text mining lie in most related research areas clustering, classification, information retrieval, machine learning and soft computing paradigms. Among all, Clustering is an unsupervised methodology having the ability to form meaningful natural groups of objects from given unlabeled data.

A large number of clustering algorithms based on K-Means have been proposed on variety of domains for different types of applications none of these algorithms is suitable for all kinds of applications. This motivated and find a room for new clustering algorithm that is more efficient and optimal with computationally feasible. Genetic algorithms are randomized optimization techniques guided by the principles of evolution and natural genetics, having a large amount of implicit parallelism. To improve the text data segmentation accuracy the authors proposed An Optimal Unsupervised Text Data Segmentation Model using Genetic Algorithm so called OUTDSM. The encoding strategy, fitness function and operators of proposed OUTDSM works together and achieve high accuracy rated optimal clusters. Additionally, the nature of biological diversity of OUTDSM prevents the population from stagnating at any local optima and promises to arrive at global optima. The experimental results proving this claim are given in this paper.

Keywords-Text Mining; Clustering; Genetic Algorithm; Optimization techniques;

1. INTRODUCTION

In today's world, the easier accessibility of digital content and the increasing capabilities of computer with user friendliness software tools generate ever increasing quantities of data in academic, economic and social activities. These activities spread the root

of text mining in customer segmentation, pattern recognition, biological studies, spatial data analysis, web document classification etc. The wide ocean of data and variety of applications has created a great demand towards text mining research to extract interesting relationships and knowledge from the text documents. Text mining has been around in many forms and has evolved greatly in its procedures and its objectives. However, extracting new knowledge from such a complex and huge amount of text documents efficiently and effectively is becoming a typical and highly important task.

Knowledge Discovery from Text Data (KDTD) is a process defined by several processing steps in order to extract potentially useful, previously unknown and ultimately understandable patterns from large amount of textual data. The important steps of KDTD are, Text Preprocessing, Text Transformation, Attribute Selection, Text Mining, Pattern Evolution and knowledge representation shown in figure 1. These steps have to be executed iteratively and several steps generally require interactive feedback from a user.

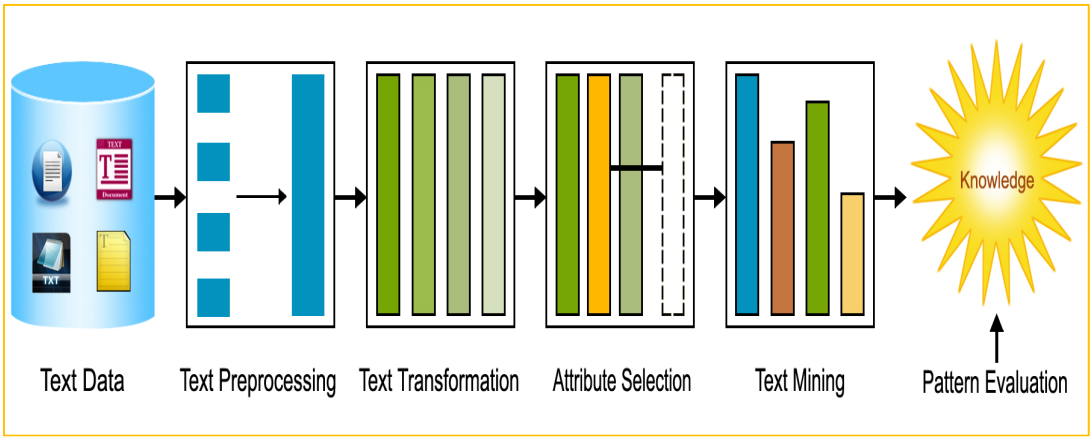

Figure 1 Steps of KDTD process

Text mining is an automated process to identify the relationships within and across enormous quantities of unstructured machine readable documents, either directly in the form of knowledge, or indirectly as functions in a fixed domain. There are several methods used to structure the text documents. Principally, these methods are classified into supervised methods which perform the text mining task by assigning a given keyword to the documents and unsupervised methods which automatically groups the similar text documents.

Indeed, it is well-known that out of all unsupervised methods, clustering is more beneficial for many real time applications with large amount of dynamically growing electronic data. Clustering is a process whose goal is to determine a finite group of documents according to some similarity concept, such that documents of the same group are as similar as possible, while documents of different groups are dissimilar. The simple usage of clustering methods in text mining is discouraged particularly due to its unsupervised nature which implies that the structural characteristics of the documents are not known in advance. This situation creates an eventual need of employing optimization techniques to provide optimal text data segmentation, and became a solid motivation for the proposed work.

Optimization techniques are stochastic evolutionary algorithms based on the principles of natural selection and natural genetics. These techniques are widely believed to

be effective on complex and non linear problems, being able to provide optimal solutions in reasonable time. Optimization techniques include genetic algorithms, evolution strategies and evolutionary programming.

In this context, Genetic Algorithms are good candidates in finding optimal and nearly optimal solutions, gradually. Genetic Algorithms are designed based on biologically inspired technology with granular computing nature and having a large amount of implicit parallelism. Their robustness and domain-independent nature motivated their applications in various fields like pattern recognition, image processing, bioinformatics, neural network design and others. Accordingly, in order to segment the text documents in an optimal way, the proposed work presents an optimal unsupervised text data segmentation model using genetic algorithm.

The remainder of this paper is organized as follows. In section 2, related work is described. In section 3, proposed work is presented in detail. In the subsequent section 4, the experimental analysis of the proposed work is shown. Finally in section 5 conclusions are mentioned.

2. RELATED WORK

The proposed work presents a brief literature survey on various advancements and techniques of text mining. It also examines how the clustering techniques are accompanied over huge electronic media data and explores the literature about optimization techniques and the usage of genetic algorithms suitable to text mining.

In the year 2007, Anna Stavrianou et.al [14] conducted a survey on semantic issues of text mining along with its approaches and methodologies proposed in the literature. They also explained the basic issues of text mining and reasons to make text mining more significant in recent research. Additionally, they covers syntactic matters, tokenization concerns, different text representation techniques, categorization tasks and similarity measures in text mining. Finally, open issues and challenges of text mining are marked in their conclusion. In the same period, M. Castellano [16] applied the text mining operations in the web environment as the web data is in semi structured or unstructured form. They designed a flexible architecture by considering the life cycle of text mining, that discover knowledge in a distributed and heterogeneous web environment. In addition, they conducted several experiments to prove the efficiency of their proposed system and results are emphasized the importance of concentrating on all the phases in knowledge discovery process in text.

In the year 2008, Milos Radovanovic, Mirjana Ivanovic [13] explored the techniques employed in text representation and several approaches to the identification of patterns in the text data. The experimental results revealed the usefulness of dimensionality reduction techniques in text mining along with clustering and classification techniques. Anna Huang [12] have compared and analyzed the efficiency of various similarity measures in text document clustering, further, they recognized that the partitioned clustering algorithms are better suited for handling large number of documents. They extended their experimental analysis and found that the performance of the cosine similarity measure is significantly better than other measures. They also expressed that the importance of selecting a similarity measure as directly impacts the performance of clustering algorithm and combined usage of these measures treated as their future work.

In 2009, Mahesh T R et.al [11] have designed a text mining framework that can apply to unstructured text documents and deduce the useful patterns. Also, they surveyed state-of-the-art text mining products and align them based on the knowledge distillation functions. They deliberately expressed that enhanced development of most efficient and scalable analysis is required in text mining. Kuan C. Chen [10] analyzed various consumer feedback forums using text mining software. They successfully identified the relationships and patterns among keywords in clusters to gain a deeper understanding of the data. They also conducted a case study to assess the effectiveness of text mining process. The methodology for semantic weight is a promising research as marked down in their future work.

In the next year 2010, Vidhya K A & G. Aghila [9] provided a detailed survey of work done so far on techniques, applications and tools of text mining. In addition, they provided in depth analysis of classification technique along with its advantages and disadvantages. They clearly mentioned that the task of text mining on unstructured text remain the largest readily available area of research. Dharmendra K Roy, Lokesh K Sharma [8] proposed a clustering algorithm on genetic k-means paradigm for mixed numerical and categorical datasets. They conducted several experiments on benchmark datasets in order to find the performance of their proposed algorithm. In their implementation process, specifically, pronounced that the importance of redesign in initialization phase and the operators of genetic algorithm.

All to range in 2011, N. El-Bathy,et.al [7] analyzed the problem of clustering using k-means algorithm and proposed an intelligent extended clustering genetic algorithm for information retrieval. Their proposed work uses several mutation operators simultaneously to get optimal solution for data clustering. Empirical testing has been performed to compare the traditional k-means with their proposed work, and the results are justified the need and relevance of genetic algorithm in document clustering process. K.Arun Prabha a, R.Saranya [6] have attempted a new context to improve the cluster quality from k-means clustering using genetic algorithm. They presented the k-means, kernal k-means and their genetic based refinement procedure in detail. They have conducted the several experiments on real datasets in medical domains and the results are evident that the genetic algorithm improves the quality of cluster analysis.

During 2012, Divya Nasa [4] presented a overall framework of text mining with its process architecture. They also discussed the earlier techniques in brief along with their pros & cons and expressed the necessity of novel approaches. Additionally, they list out the benefits and limitations of text mining. In the same period, Dr. A.V. Senthil Kumar, S.Mythili [5] proposed a new algorithm for parallel implementation of genetic algorithm using k-means clustering. They conducted the experimental analysis on artificial datasets and found the importance of genetic algorithm process in finding the good clustering configuration.

Recently in 2013, Rashmi Agrawal, Mridula Batra [2] have studied the concept of text mining and various existing techniques. They also described the major ways in which text is mined when the input is only plain or structured text. Further, they expressed that once the preprocessing of the document has been completed, various well known analytical techniques such as clustering, factoring can be used for further processing. In the same period, Deepankar Bharadwaj [1] have implemented classification and prediction methodology as a two-step procedure which can mine the details from text resumes and give the optimized

solution on the basis of the information extraction. They concluded that the combination of text mining and genetic approach is relevant area of research.

Many of the earlier authors in the literature have clarified the significance and criticality of genetic approach in the clustering process of text data mining, which has been considered as the fundamental basis for the proposed work that encourages the authors to describe the problem.

3. PROPOSED OPTIMAL UNSUPERVISED TEXT DATA SEGMENTATION MODEL(OUTDSM)

The efficiency and quality of text data segmentation by clustering techniques must be examined whether this segmentation process is optimally identifying the similarity between the text documents. To improve the accuracy of text data segmentation, usage optimization techniques is becomes one of the central problems in the segmentation process. With these objectives the authors propose an Optimal Unsupervised Text Data Segmentation Model (OUTDSM) as shown in figure 2 with its architecture. The OUTDSM, initially, prepares the structured data efficiently by designating a vector space model with its procedural steps and filter methods. In the next stage, the proposed model emphasizes on identifying the initial segments by employing a standard k-means algorithm with its Euclidian distance approach. Finally, the OUTDSM concentrates on generation of optimal segments using Genetic Algorithm with its stochastic nature.

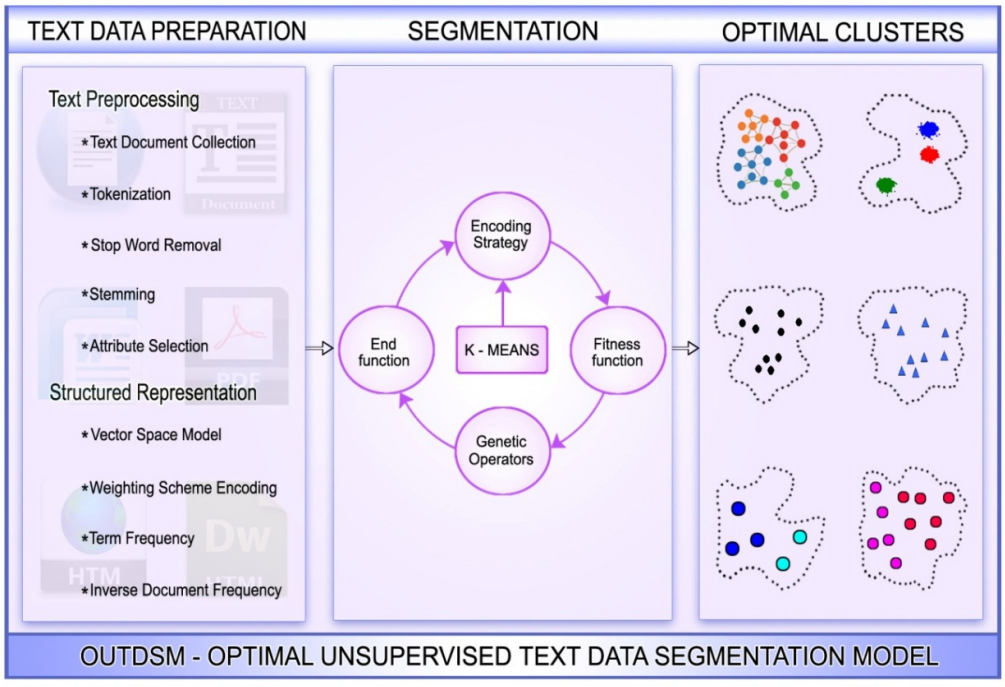

Figure 2 Architecture of proposed OUTDSM

3.1 TEXT DATA PREPROCESSING

The preprocessing is the crucial task in KDTD process and in fact, the effective and successful text mining operations are highly dependent on sophisticated preprocessing methodologies that extract structured representations from unstructured text documents. Towards this, the first task of OUTDSM, concentrates on preprocessing of text documents, basically, aims at transforming the text collection into a useful form for text mining algorithms. The importance of preprocessing is emphasized by the fact that the quality of categorization of documents. The text preprocessing of OUTDSM covers all the procedural steps including text document collection, tokenization, stop word removal, stemming and attribute selection.

Text Documents Collection: The first step of text data preprocessing of OUTDSM is the collection of text documents. The text collection process totally depends on the goal of the text mining over the digital universe.

Tokenization: In this step, the proposed system explores the words as tokens by splitting a text into pieces and removing all punctuation marks by parser. These token representations are separated by replacing a single white space between the words as word boundaries and then used for further processing. The set of words or tokens obtained by merging all text documents of a collection is treated as the dictionary of OUTDSM. The OUTDSM parser caters the consistency for different number and time formats in the documents. Moreover, it identifies meaningful keywords and transforms the abbreviations and acronyms into a standard form. OUTDSM successfully handles the phase of a tokenization and it will allow for an accurate analysis of the document during the Text Mining.

Stop Word Removal: Stop words are words with little or no semantic value as they are used to provide structure in the language rather than content. In general, the stop word list consists of pronouns, prepositions, conjunctions etc. which are language-specific functional words, are frequent words that carry no information. Yet, these words may also confuse the entire text mining operations. Moreover, the removal of these worthless words reduces the dimensionality of the vector space model. In order to get the benefits from removing the stop words, in the next step, OUTDSM eliminate the most frequently used words in English that are useless for text mining. In addition, the plain stop words such as names of day and month, terms with one or two characters and all terms containing non alphabetical characters are carefully discarded during this step. Moreover, the template words, means the terms printed on each page of a document are also deleted as they might influence results.

Stemming: In order to further reduce the size of the vector space model and increase the performance of learning algorithm, the OUTDSM take up the task of stemming. Stemming is a process to reduce different grammatical or word forms of word to its consolidated root or stem form as meaning is carried predominantly by its root. The goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common-base form. The common word forms are noun, adjective, verb, adverb, plurals etc. This procedure incorporates a great deal of language dependent linguistic knowledge but not on the domain. The stemming process majorly considers two points: one is the morphological form of a word that has the same base meaning should be mapped to the same stem. The other is the words

that do not have the same meaning treat as separate. To do this, the OUTDSM employs a well known statistical N-Gram stemmer, which uses string similarity approach to covert word inflation to its stem. The main idea behind this approach is that, similar words will have a high proportion of n-grams in common. This process is useful in the clustering as it makes the operation less dependent on particular forms of words.

Attribute Selection: Even after a thorough process of cleaning the text documents, the set of attributes is still huge. In order to reduce further the size of attribute set, the OUTDSM moves towards attribute selection as another pre processing step. Attribute selection is a technique which selects better attributes to delimitate the problem domain and, consequently, contributing to improve the performance of the learning algorithms used in the knowledge extraction step. Towards this, in this step, the OUTDSM computes the following main measures as filters based on variance representativeness.

- **Term Contribution:** Term contribution is calculated on the basis of how much one specific term contributes to the similarity among all collected text documents.
- **Term variance:** This measure finds the variance of all terms in all the text documents. Then, assign the highest scores to terms that have high document frequency and a uniform distribution through the text documents collection.
- **Term Variance Quality:** This measure calculates the quality of a term by using the total variance.

Finally, these filters assign ranks to all the attributes, so the top-ranked attributes are taken for further processing.

Text Data Encoding and Structured Representation using Vector Space Model:
To arrange the preprocessed text data into a suitable format for text mining techniques, it is necessary to use document encoding methodology. The simplest way of document encoding is to use Boolean binary representation; an element is set to '1' if the corresponding word is used in the document and to '0' if the word is not. The simple binary encoding gives similar significance to many documents. To overcome this, OUTDSM employs term weighting scheme encoding, where the weight reflects the importance of a word in a specific document. Large weights are assigned to terms that are used frequently. The document weight w(d, t) for a term 't' in a document 'd' is computed by the product of term frequency TF(d, t) with inverse document frequency IDF(t). Based on document weight, the OUTDSM prepares vector space model that represents document as vectors in m-dimensional space. Here each document d is described by a numerical feature vector $W_d = \{x(d, t_1), x(d,t_2), \ldots\ldots x(d, t_m)\}$. Now, the documents can be compared by performing simple vector operations for text mining process. A sample dimensional vector space model is as shown in table 1.

Table 1: A Sample Dimensional Vector Space Model

	Team	Play	Coach	Ball	Game	Score	Win	Lost	Timeout	season
Doc 1	3	0	5	0	2	6	0	2	0	2
Doc 2	0	7	0	2	1	0	0	3	0	0
Doc 3	0	1	0	0	1	2	2	0	3	0

3.2 OUTDSM TEXT DATA SEGMENTATION USING STANDARD K-MEANS

The structured text data generated by vector space model is the basic input for the k-mean clustering algorithm. All documents in vector space model are partitioned into the disjoint subsets. In k-means algorithm, k initial centroids are chosen where k is the user specified number of clusters. Each text document in the vector space model is assigned to the closest centroid and each collection of text documents assigned to a centroid is a cluster. The centroid of each cluster is then updated based on text documents assigned to the cluster. The process of assignment and updating of cluster is replaced until the centroids remain the same. The k-means is formally described by the following algorithm.

Algorithm 1: k-Means Clustering Algorithm:

Step 1: Select k text documents as initial centroids
Step 2: Repeat
Step 3: From k-clusters assigning each text document to its closest centroid
Step 4: Recompute the centroid of each cluster
Step 5: Until centroids do not change.

In the above algorithm the assignment and reassignment of a document to the cluster is done on the basis of Euclidian distance. The computation and re-computation of centroid is calculated with the help of statistical mean. Here, the quality of the clustering is estimated using Sum of Squared Error (SSE). Thus, the standard k-mean algorithm guaranteed to find only local optimal solution. Further, selection of poor initial centroids and random initialization of k-mean algorithm leads to low accuracy in the clustering process.

3.3 OUTDSM OPTIMAL TEXT DATA SEGMENTATION USING GENETIC ALGORITHM

While identifying the cluster, the standard k-mean clustering technique has to take many decisions at different levels. The standard k-mean algorithm impacted by many external and explicit conditions like selection of poor initial centroids and random initialization, the clustering process becomes a non linear problem. Thus, the conventional clustering techniques used for this type of problem happen to computationally expensive and leads to only local optimal solution. To attain global optimal clusters, the authors propose a genetic approach clustering technique in OUTDSM.

Optimization is a method of finding the global solution under given conditions. There are many optimization techniques in the literature, out of which Genetic algorithm is well suited to find the optimal clusters. To get the optimal and quality clusters, it is necessary to process the clusters generated by standard k-mean algorithm.

The proposed OUTDSM considers each step of genetic algorithm equally with the goal of finding the optimal clusters. The first step of genetic algorithm considers k clusters generated by k-Means algorithm as its initial population. Then, the biological genetic operators create a new and potentially better population. Later, by the theory of evaluation only optimal individuals survive based on fitness function. Finally, this process is terminated as and when an acceptable optimal set of clusters is found.

OUTDSM Encoding Strategy-Initial Population: The encoding strategy is a process of representing the potential solution to a problem into a suitable form of viable individuals so that the genetic algorithm can process. It is a crucial issue in the genetic process as it plays a critical role to arrive at best performance of algorithm as robust as possible. Various encoding strategies have been introduced in the literature for effective implementation of genetic algorithm. OUTDSM adopts the Binary encoding strategy to determine initial population from the clusters generated by k-Mean algorithm. Consider following example of cluster {D1, D4, D5, D8} is encoded as a binary chromosome of length 10 and is shown in figure 3. The presence of a document in a cluster is coded as 1, otherwise as 0.

Figure 3. Example Binary Chromosome

OUTDSM-Fitness Function: The characteristics of fitness function play a key role in implementing a successful genetic algorithm. In addition, the convergence is highly depends on the number of iterations for which the maximum fitness function value occurs. The goal of genetic algorithm is typically expressed by its fitness function that evaluates rate of optimality in the resultant clusters. The OUTDSM fitness computation process is performed with the combination of two important measures. One is cluster cohesion, which determines how closely related documents in a cluster. Another is cluster separation, which determines how distinct a cluster from other clusters.

The cohesion of a cluster is calculated by the sum of the proximities with respect to the prototype of the cluster. Similarly, the separation between two clusters is measured by the proximity of the two cluster prototypes.

$$cohesion(C_i) = \sum_{X \in C_i} Proximity(x, c_i) \quad (1)$$

$$separation(C_i, C_j) = promity(c_i, c_j) \quad (2)$$

Where c_i is the prototype (centroid) of cluster C_i and c is the overall prototype (centriod). Proximity function is function is a similarity or dissimilarity measure. In this context, OUTDSM uses the cosine similarity measure to find the value of proximity function. Cosine similarity is defined as follows,

If x and y are two documents vectors, then

$$cos(x, y) = \frac{x.y}{\|x\|\|y\|} \quad (3)$$

Where $x.y = \sum_{k=1}^{n} x_k y_k$, and $\|x\|$ is the length of vector x, $\|x\| = \sqrt{x.x}$

OUTDSM-Operators: Biologically inspired operators like selection, crossover and mutation are applied on the clusters to yield a new generation of clusters. The process of selection, crossover and mutation continues for a fixed number of generations or till a termination condition is satisfied.

Selection: The selection process selects chromosomes from the mating pool directed by the survival of the fittest concept of natural genetic systems Here the OUTDSM deploys the roulette wheel parent selection procedure. This wheel as many slots as population size where the size of the slot is proportional to the relative fitness of corresponding cluster chromosome in the initial population as demonstrated in figure 4. An individual cluster is selected by spinning the roulette and noting the position of the marker when the roulette stops. Thus, the number of times the selection of individual cluster is proportional to its fitness function value in the population.

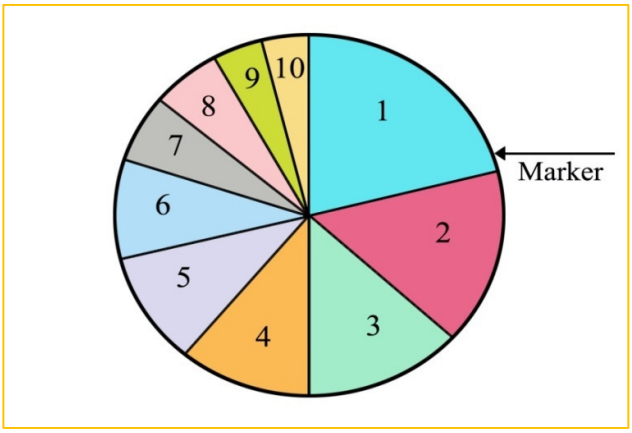

Figure 4 Example of Roulette Wheel Parent Selection

Crossover: Crossover is a probabilistic process that exchanges information between two randomly selected parent chromosomes for generating two child chromosomes. In this paper, single point crossover with a fixed crossover probability of Cp is used. For chromosomes of length l, a random integer, called the crossover point, is generated in the range [1, l-1]. The portions of the chromosomes lying to the right of the crossover point are exchanged to produce two offsprings. An example is as shown in figure 5 and 6 before and after crossover respectively.

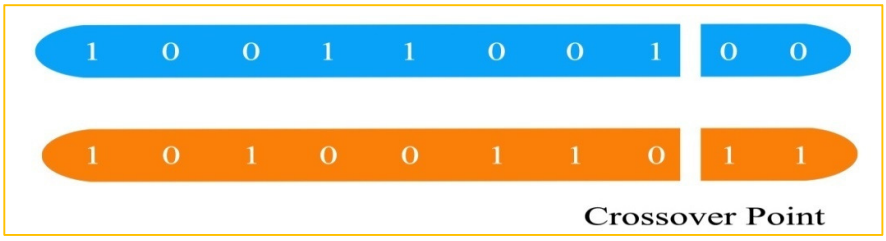

Figure 5 Single Point Crossover Operations before Crossover

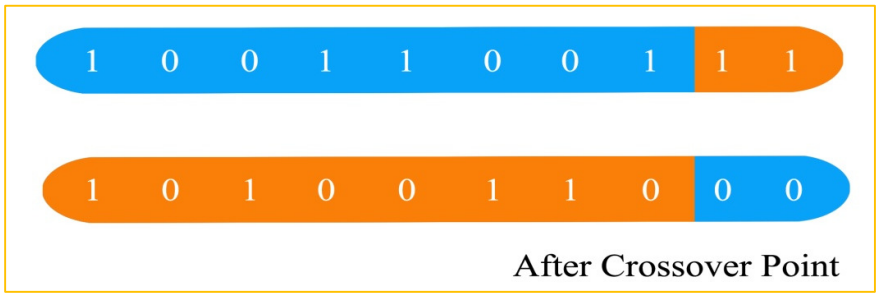

Figure 6 Single Point Crossover Operations after Crossover

Mutation: Each chromosome undergoes mutation with a fixed probability Mp. For binary representation of chromosomes, a bit position is mutated by simply flipping its value. Since we are considering binary values in this paper, a random position is chosen in the chromosome and replace by negation of bit. An example is shown in figure 7.

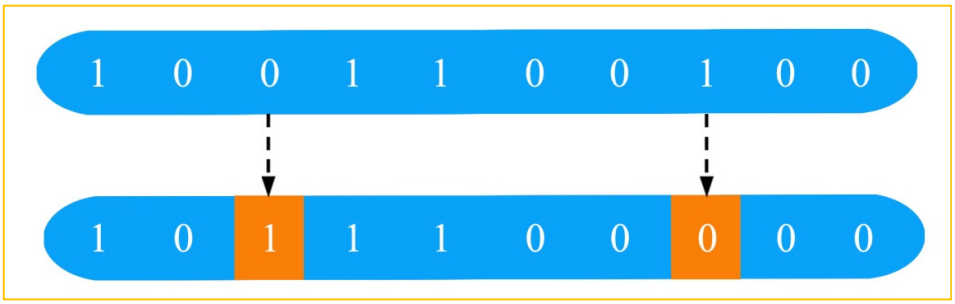

Figure 7 Process of bit-by-bit mutation

OUTDSM-Genetic Algorithm:

Step 01: Start
Step 02: Prepare initial population using Encoding Strategy
Step 03: Evaluate Initial population using Fitness Function
Step 04: Generate New Population by Repeating
Step 05: Apply Selection Operator
Step 06: Apply the crossover Operator
Step 07: Apply the Mutation Operator
Step 08: Until termination condition is met (No change / reach user threshold)
Step 09: Return optimal data segments
Step 10: Stop

4. EXPERIMENTAL ANALYSIS

The proposed OUTDSM has been implemented in a standard environment. For the OUTDSM genetic algorithm, number cluster generated by standard k-Mean algorithm are given as input. Theses clusters are generated by standard k-Mean algorithm from structured data from represented in vector space model consists of more than 3000 documents.

A. Several experiments are carried out to demonstrate the performance of OUTDSM with respective accuracy in text data segmentation. It is observed from figure 8 that the OUTDSM significantly performs with high rate of accuracy on number of iterations.

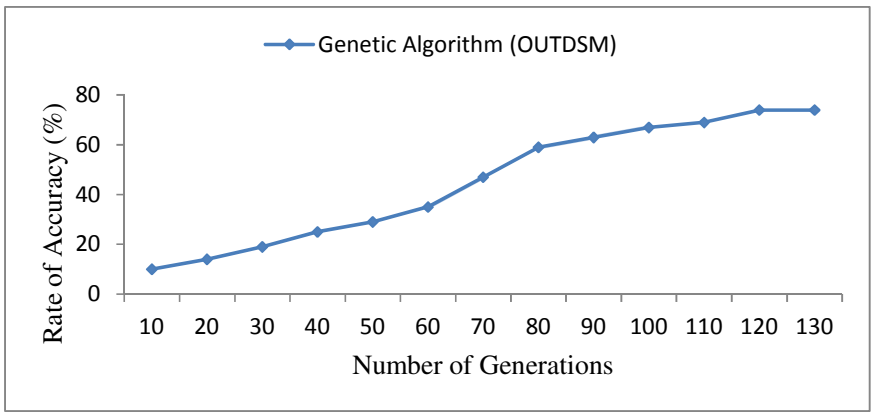

Figure 8 Accuracy rate of OUTDSM

B. The OUTDSM experimented for different cross over probabilities (Cp=0, 0.25, 0.5, 0.75, 1) for different number of iterations. It is showing high performance at the average mean Cp as in Figure 9.

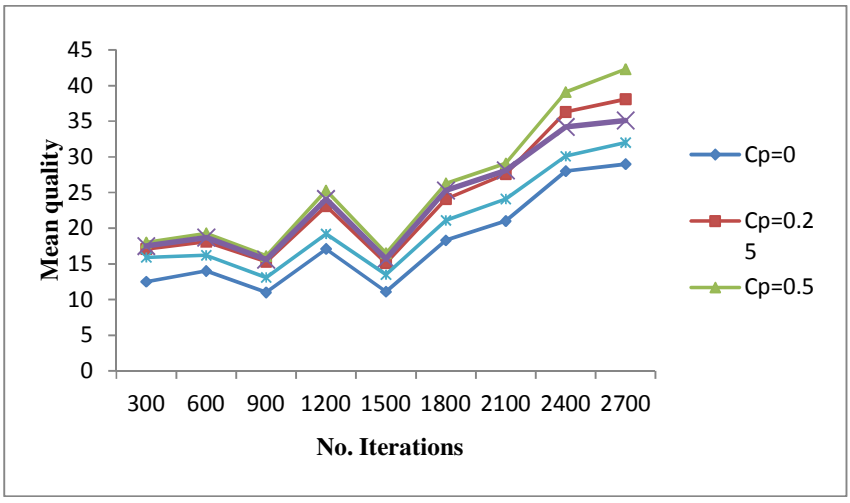

Figure 9 Population averaged Mean quality of OUTDSM

C. The OUTDSM compared with the standard k-Mean clustering technique with respective Execution. The experimental result shows the OUTDSM is essentially taking more time when compared with k-Mean clustering technique as shown in figure 10, in generating optimal clusters.

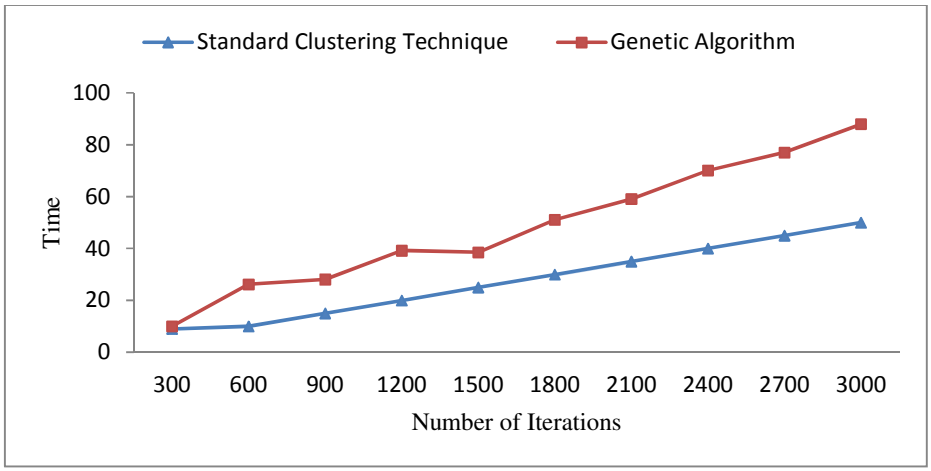

Figure 10 Accuracy rate of OUTDSM

5. CONCLUSION

The present work OUTDSM is designed with genetic approach and it takes the input from a standard k-Mean clustering technique for optimally segments the text data. These optimal data segments certainly help to improve the business of any organizations by analyzing, categorizing and in drawing conclusions about their customers and competitors. The experimental results are evident that the proposed OUTDSM segment the text documents intelligently.

ACKNOWLEDGEMENTS: The authors would like to thank the Department of Science & Technology (DST), Ministry of Science & Technology, Government of India under Women Scientist Scheme A (WOS-A) for providing the fund to this research. The authors also recorded their acknowledgements to the authorities of Shri Vishnu Engineering College for Women, Bhimavaram, A.P., India for their constant support and cooperation.

6. REFERENCES

1. Deepankar Bharadwaj, "Text Mining Technique using Genetic Algorithm", International Conference on Advances in Computer Application (ICACA - 2013), pp:7-10, 2013.
2. Rashmi Agrawal, Mridula Batra, "A Detailed Study on Text Mining Techniques", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.

3. V.V.R. Maheswara Rao, Dr. V. Valli Kumari, "An Intelligent Optimal Genetic Model to Investigate the User Usage Behaviour on World Wide Web", International Journal of Data Mining & Knowledge Management Process Vol.3, No.2, 2013.
4. Divya Nasa, "Text Mining Techniques- A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, 2012.
5. Dr. A.V. Senthil Kumar, S.Mythili, "Parallel Implementation of Genetic Algorithm using K-Means Clustering", Int. J. Advanced Networking and Applications, Volume:03 Issue:06 Pages:1450-1455, 2012.
6. K.Arun Prabha a, R.Saranya, "Refinement of K-Means Clustering Using Genetic Algorithm", Journal of Computer Applications (JCA) ISSN: 0974-1925, Volume IV, Issue 2, pp: 40-44, 2011.
7. N. El-Bathy, C. Gloster, I. Kateeb, G. Stein, "Intelligent Extended Clustering Genetic Algorithm for Information Retrieval Using BPEL", American Journal of Intelligent Systems, Vol 1(1): pp: 10-15, 2011.
8. Dharmendra K Roy and Lokesh K Sharma, " Genetic k-Means Clustering Algorithm for Mixed Numeric and Categoorical Data Sets", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.2, pp:23-28, 2010.
9. Vidhya. K. A & G. Aghila," Text Mining Process, Techniques and Tools : An Overview", International Journal of Information Technology and Knowledge Management, Volume 2, No. 2, pp. 613-622, 2010.
10. Kuan C. Chen, "Text Mining e-Complaints Data from e-Auction Store with Implications for Internet Marketing Research", Journal of Business & Economics Research, Volume 7, Number 5, pp: 15-24, 2009.
11. Mahesh T R, Suresh M B, M Vinayababu, "Text Mining: Advancements, Challenges and Future Directions", International Journal of Reviews in Computing, pp: 61-65, ISSN: 2076-3328, 2009.
12. Anna Huang, "Similarity Measures for Text Document Clustering", New Zealand Computer Science Research Student Conference, pp:49-56, 2008.
13. Milos Radovanovic, Mirjana Ivanovic "Text mining: Approaches and Applications", Novi Sad J. Math, Vol. 38, No. 3, pp: 227-234, 2008.
14. Anna Stavrianou, Periklis Andritsos, Nicolas Nicoloyannis, "Overview and Semantic Issues of Text Mining", SIGMOD, Vol. 36, No. 3, pp: 23-34, 2007.
15. Lipika Dey, Muhammad Abulaish, Jahiruddin, Gaurav Sharma, "Text Mining through Entity-Relationship Based Information Extraction", IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology – Workshops, pp:177-180, 2007.
16. M. Castellano, G. Mastronardi, A. Aprile, and G. Tarricone, "A Web Text Mining Flexible Architecture", World Academy of Science, Engineering and Technology, pp: 78-85, 2007.
17. Elizabeth Leon, Olfa Nasraoui, Jonatan Gomez, "ECSAGO: Evolutionary Clustering with Self Adaptive Genetic Operators", IEEE Congress on Evolutionary Computation Sheraton Vancouver Wall Centre Hotel, pp: 1768-1175, 2006.
18. Louis A., Francis, FCAS, MAAA, "Taming Text: An introduction to Text mining", Casualty Actuarial Society Forum, pp: 51-88, 2006.
19. Joel D. Martin, "Fast and Furious Text Mining", and IEEE Computer Society Technical Committee on Data Engineering, pp: 1-10, 2005.