DATA LEAK IDENTIFICATION IN SOCIAL NETWORKS USING HYBRID TRANQUILITY K MEANS CLUSTERING

Jayavarapu Karthik
Research Scholar, Department of Computer Science & Engineering, Annamalai University,
Chidambaram, Tamilnadu-608002, India.
Jayavarapukarthik@gmail.com.

V. Tamizhazhagan
Assistant professor, Department of Information Technology, Annamalai University,
Chidambaram, Tamilnadu-608002, India.
rvtamizh@gmail.com.

Satyala Narayana
Professor, Department of Computer Science & Engineering, Seshadri Rao Gudlavalleru Engineering College,
Gudlavalleru, Andhra Pradesh-521356, India.
Satyala1976@gmail.com.

Abstract

Loss of sensitive data can be stopped employing Data Leak Prevention (DLP). Most of such tools can be quite effective while protecting private information known already. At the same time, plenty of private information has not been recognized until it has been disclosed to various unknown users or other competition enterprises. Clustering refers to a data mining technique that can classify a certain set of instances into different clustersusing a measure of similarity. One of the most common algorithms based on partitioning is the K-Means. However, it has many drawbacks like it can generate local optimal solutions that are based on initial centroids that are chosen randomly. The Tabu Search (TS) Tranquility Search and the Stochastic Diffusion Search (SDS) have been proposed in this work. In one of the most recent algorithms called Tranquility Search, optimal global solutions are obtained by exploring through the entire solution space. Some studies show hybrid algorithms that are a combination of two different ideas producing better solutions. In this work, a new approach is presented which is a combination of two different ideas producing better solutions. The Improved Tranquility Search technique and the K-means algorithm are combined. For this, a hybrid Tranquility-TABU-SDS algorithm is applied in the social network for the DLP. The results of the experiment have proved that the method proposed performs better in comparison to other methods.

Keywords: **Social Network; Data Leakage Prevention (DLP); K-Means Clustering, Tabu Search (TS); Stochastic Diffusion Search (SDS); Tranquility Search Algorithm**.

1. Introduction

In social media, thecommunication among the data owners and the viewer or the end-user creates virtual communities with Online Social Networks (OSN). The relation between the organizations and their users along with their various social activitiesis represented as a social graph. Such users, groups, and organizations will be the edges of a graph. The OSN will be an online platform that is used by various end-users in creating relationships and social networks with other people sharing similar views, activities, and interests [1].

The primary goal of the OSN was to share content with the maximum number of users. The users will make use of the OSNs like LinkedIn, Twitter, and Facebook in order to publish some of their routine activities. At times, the OSN users may also share information relating to their lives and also the lives of their colleagues and friends. But, in the case of published data, some contents that are exposed by the OSN can be private and thus must not be published. Ideally, the users will share only partly their lives using status updates, videos, or

photographs. Currently, there are different OSN users that make use of smartphones for taking pictures or for making videos to be shared through the OSNs. Such data may have information on location along with some metadata that is embedded in it. The providers of OSN services may collect a wide range of data with regard to users in order to offer certain personalized services, and this may also be utilized for commercial goals. Additionally, the user data can be given to third parties that may result in leakage of privacy. Such information may permit malicious users to invade individual privacy [2].

Data Leak Prevention (DLP) has been identified as a field of research that deals with the investigation of certain potential threats to security in the organizational strategies and data in order to mitigate these attacks. Data leaks may include the release of certain sensitive information to third parties either intentionally or otherwise as the attacker or hacker gains access to sensitive data. Loss of data, however, refers to damage or disappearance of data where the right copy of data is not available any longer, thus compromising the availability and integrity of such data. There are several DLP products and techniques that are available attempting to mitigate these threats like McAfee. The DLP is a problem that is yet to be solved as the current products continue to remain limited to the threats. Recently there has been some development to the increase in the attacks on the organizations resulting in major concern to the entire globe. There are many Cyber-attack reports that are often making headlines [3]. Data leak identification can help prevent the leaks and the mechanism for data leak identification in social networks is shown in figure 1.

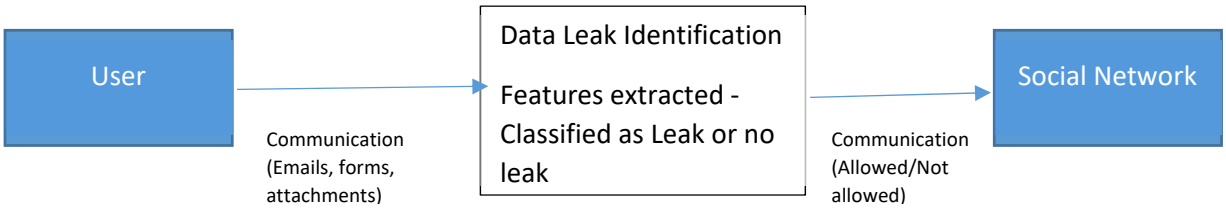

Figure 1: Data Leak Identification.

A data leak can be a frequent one in an organization until it remains undetected. Loss of data can, however, lower. Data loss and data leak may not always be malicious. Sometimes, there may be incidents of natural disasters that can destroy physical structures. There may also be careless handling of printed documents that are the sensitive or careless entry of data, both being unintentional. Traditional forms of data leak or data loss owing to carelessness, virus attacks, and natural disasters will need the DLP approach. This can be in the form of waterproof/fireproof cabinets, antimalware or antivirus protection, watermarking, and access control. But sophisticated data leak needs a specialized solution [4].

Clustering can be a problem of data mining in text domains. There are several applications that are made towards customer classification, segmentation, indexing, document organization, visualization, and collaborative filtering. It refers to the classification of various objects into groups. The partitioning of datasets into clusters in order to ensure each subset is able to share a common trait in accordance with a certain level of distance measuring is used here. Clustering is further classified into different types, which are grid-based, constraint-based, density-based, co-clustering, relocation clustering, and partitioning. K-Means Clustering is a new method of partition that has mutually exclusive clusters that are spherical in shape. It can generate a certain number of flat (non-hierarchical) and disjoint clusters that organize the objects into k –partitions wherein every partition is a cluster [5].

Even though the K-means algorithm is simple, straightforward, and easy to implement, it does suffer from some drawbacks rendering it unsuitable for some applications. The main disadvantage is defining of the number of K clusters even before its application. Further, the summary statistic being the mean of its values for every cluster, the individual members in each cluster will need to have a high variance without a good summary. Also, with the increase in the number of clusters, it can become untenable, and this will approach the O (n^2) comparisons when n becomes the actual number of instances. One more major issue with the K-means algorithm can be its sensitivity to initialization. It may converge to the local optima and also the final clusters that are not necessarily be optimal [6]. For overcoming the problem of local optima, plenty of investigation has been made. Here, researchers make use of meta-heuristic algorithms such as the, Genetic Algorithm (GA),TABU Search (TS), Ant Colony Optimization (ACO).

Meta-heuristic algorithms mimic the process of improvisation of music players developed recently called tranquillity search. This is very successful in various problems of optimization, presenting many advantages in connection to the other traditional techniques of optimization. The features further increase the tranquillity and flexibility of the algorithm resulting in better efficiency. As these stochastic optimization approaches are great in avoiding convergence, they may be used to identify globally optimal solutions that ideally need a longer time frame [7]. Here, the hybrid Tranquility-TABU-SDS algorithm was proposed for social network DLP. The remainder of the investigation is presented thus: Section 2 talks about some of the works available in the literature. Section 3 explains the techniques used. Section 4 presents the results of the experiments, and the conclusion is given in Section 5.

2. Related Works

Zilberman et al. [8] presented a technique for detecting data leak through emails. The emails exchanged among the members of the organization is analysed and based on the information, group of members and common topics are identified. When a new mail is sent, it checks for recipients and content of the mail. If it is an approved recipient and common topic the mail is sent. Fan and Wang [9] had made a proposal of a mean-field method that is used for spreading fake news based on PSO-based networks that assumes that unawareusers tend to accept as true and report fake news. Thus, the rate of spreading has been related to any similarity among the individuals. The Monte-Carlo simulations implemented based on the proposed method demonstrates the efficiency. Kaur et al. [10] had made an attempt to survey various techniques for the prevention of data leakage, aside from certain challenges and data protection approaches, and also accounted for the constraints. Academics and professionals, we're able to draw benefits from the survey.Botti-Cebriá et al., [11]provided the sensitivity of the data that the user would share when published online. To this end, an assistant agent was proposed for the identification of sensitive data that was dependent on the message's distinct types of detected categories (that is, emotions, personal attacks, health, personal data, location, and so on). The distinct categories were detected through the utilization of sentiment analysis, dictionaries, ontologies, and entity recognition libraries.

Gupta and Kush [12] had concentrated on the data leakage, its methods as well as the DLD modules for detection. In addition to that, there was the presentation of a literary review of the various data leakage methods. A watermarking method is utilized for dealing with the data leakage. This, in turn, would result in the data's alteration. In the event that the altered watermark data copy's location was at an unallowed site, the distributor could allege that he had rights there. SocialCrowd, a data leak-aware crowdsourcing system, was recommended by Amor et al., [13]. This system introduced a clustering algorithm that used the crowd workers' social relationships for the discovery of all potential teams whilst avoiding inter-team data leakage. This system had also defined a ranking mechanism for the selection of the "best" team configurations.For intentionaland accidentalDLP, Katz et al., [14] had propounded a novel context-based method (CoBAn). For the prevention of data leakage, existing methods tried to either seek specific phrases and keywords or to utilize numerous statistical techniques. Since the keyword's context would get ignored, keyword-based methods were not satisfactorily accurate. On the other hand, the analyzed text's content would get ignored in statistical methods. The proposed context-based model would leverage the benefits of both earlier methods and would be composed of two stages: the training stage and the detection stage. In the training stage, there would be the generation of clusters of documents as well as the creation of a graph representation of each cluster's confidential content. The graph representation would be made up of key terms as well as the context in which they must appear so as to be considered confidential. In the detection phase, there would be an allocation of each tested document to various clusters. Later on, its contents would get matched to each cluster's associated graph for the determination of the document's confidentiality. In comparison to other methods, it was evident from the comprehensive experimentations that the proposed model was superior in the detection of leakage attempts, in which the confidential information either was rephrased or was different from the learning set's given original examples.

3. Methodology

One of the most commonly used algorithms today has been the K-Means, which is not just computationally efficient but greedy as well. This K-means algorithm [15] is implemented as follows: (1) Initializing of the cluster centroids: this is implemented by making use of various methods. The most commonly used one is randomly generated centroids. (2) Reiterate this until such time the convergence takes place: (a) A data point is assigned to a cluster based on its distance from the centroids. (b) updating the centroids values by calculating the average of the point attribute values that belong to the cluster. Here, Tranquility K-Means, Tranquility-TABU-SDS, and Tranquility -SDS K-Means have been explained.

The proposed algorithms were evaluated using Enron datasetis made up of personal as well as official emails. The majority of the integrity issues for this dataset were resolved. There was the deletion of some emails

due to appeals from affected employees. This specific dataset version was made up of almost 517,431 mails which were retrieved from 151 users who were spread across 3,500 folders. However, no attachments were included in these messages.This dataset had folders of information on all 151 employees. Each message in the folder had the receiver's email address, the sender's email address, time, date, subject, body, text, and a few other technical details.

3.1 *Tranquility K-Means Algorithm*

The search ability of K-Means is discussed by defining a fixed set of K cluster centres found in R^N; this ensures suitable formations of cluster of the n unlabelled points. The Euclidean distance for the points from cluster centres, based on which the following steps are proposed [16]:

1: Initializing the algorithm

One commonly used performance measure for finding the goodness of k clustersis the total mean-square quantization error (MSE), which is given for *k* number of clusters as (1):

$$Minimize\ f\ (X, C) = \sum_{n=1}^{N} Min\{||x_n - c_l||^2 | l = 1, \ldots, K\} \qquad (1)$$

Wherein *f (x, c)* refers to the objective function, $x_n$is the data, *N* refers to the number of data, $c_l$denotes the cluster center. *C* will be the set for each decision variable C_l,

The parameters of the tranquility search algorithm are Tranquility Memory Size (TMS), a number of solution vectors of the Tranquility Memory (TM); Pitch Adjusting Rate (PAR), Tranquility Memory Considering Rate (TMCR); and Number of Improvisations (NI), or the stopping criterion.TM refers to a memory location in which the solution vectors have been saved. The TMCR and the PAR are the parameters used for improving their solution vector. Both of these have been defined in Step 3.

2: Initializing of tranquility memory

Here, the TM matrix will be populated with the maximum number of randomsolution vectors called the TMS as in (2):

$$\begin{bmatrix} c_{11}^{1} & c_{12}^{1} & \ldots & c_{k(d-1)}^{1} & c_{kd}^{1} \\ c_{11}^{2} & c_{12}^{2} & \ldots & c_{k(d-1)}^{2} & c_{kd}^{2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ c_{11}^{TMS-1} & c_{12}^{TMS-1} & \ldots & c_{k(d-1)}^{TMS-1} & c_{kd}^{TMS-1} \\ c_{11}^{TMS} & c_{12}^{TMS} & \ldots & c_{k(d-1)}^{TMS} & c_{kd}^{TMS} \end{bmatrix} \qquad (2)$$

Step 3: Improvising a new tranquility

A tranquility vector $c' = (c_{11}^{'}, c_{12}^{'}, \ldots, c_{1d}^{'}, c_{21}^{'}, c_{22}^{'}, \ldots, c_{2d}^{'}, \ldots, c_{k1}^{'}, c_{k2}^{'}, \ldots, c_{kd}^{'}) = (c_{ij}^{'}) 1 \leq i \leq d\ and\ 1 \leq j \leq k$is createdusing three guidelines: (1) consideration of memory, (2) adjustment of the pitch, and (3) random selection. The generation of new tranquility is known as 'improvisation'. For memory factor, the actual value of the first decision variable $(c_{11}^{'})$ used for a new vector is selected from a value that is specified in a particular range of TM $(c_{11}^{1} - c_{11}^{TMS})$ TMS. The values for the other decision variables are $(c_{12}^{'}, \ldots, c_{1d}^{'}, \ldots., \ldots, c_{kd}^{'})$ selected in the same way. TMCR, falling between 0 and 1 will be the rate of selecting one value from prior values in the TM; (1-TMCR) refers to the rate of choosing one value from a range of values randomly as in (3).

$$c_{ij}^{'} \leftarrow \begin{cases} c_{ij}^{'} \in \{c_{ij}^{1}, c_{ij}^{2}, \ldots, c_{ij}^{TMS}\}\ with\ probability\ TMRCR, \\ c_{ij}^{'} \in c_{ij} with\ probability\ (1 - TMCR) \end{cases} \qquad (3)$$

For instance, the TMCR for 0.85 is a tranquility search algorithm that selects the value of the decision variable from values saved in the TM. This has an 85% probability or will from the whole range that has a probability of 100– 85 %. Each component from memory will be examined to check if it has to be pitch-adjusted. The operation will use the PAR parameter that is the rate for adjustment of the pitch as depicted in (4):

$$Pitch\ adjusting\ decision\ for\ c_{ij}^{'} \leftarrow \begin{cases} Yeswith\ propability\ PAR, \\ Nowith\ propability\ (1 - PAR), \end{cases} \qquad (4)$$

The value for (1 - PAR) will set the rate for doing nothing. In case the decision for pitch adjustment $c_{ij}^{'}$ is YES, $c_{ij}^{'}$ then it is replaced as in (5):

$$c_{ij}^{'} \leftarrow c_{ij}^{'} \pm rand() \times bw \quad (5)$$

Wherein *bw* is the arbitrary distance bandwidth, and the *rand* () refers to the random number falling between 0 and 1.

As in Step 3, the TM factor, pitch adjustment, and random selectionare utilized to the variables for the new tranquility vector.

Step 4: Updating the tranquility memory

If the new tranquility vector $c' = (c_{11}^{'}, c_{12}^{'}, \ldots, c_{1d}^{'},, c_{k1}^{'}, c_{k2}^{'}, \ldots, c_{kd}^{'})$was found to be better than the worst tranquility of the TM, the current one is incorporated in the TM, and the worst is omitted.

Step 5: Checking of the stopping criterion

The maximum number of improvisations is considered as the stopping criterion, if this is met, the algorithm is terminated. If not, Steps 3 and 4 will be repeated.

Step 6: Check to stop criterion

If the stopping criterion (maximum number of improvisations) is satisfied, computation is terminated. Otherwise, Steps 3 and 4 are repeated.

3.2 *Tranquility -SDS K-Means Algorithm*

Stochastic Diffusion Search (SDS) refers to a population-based algorithm implementing patterns of direct communication for evaluating the hypothesis of search and optimization. The agent population will have a 'hypothesis' with regard to the solutions, and they are evaluated partially for providing feedback to make sure the agents are in convergence with the promising solutions. By making use of the SDS, the agent communication along with the partial evaluation of the hypothesis has a major role to play in intelligence. The SDS algorithm has three phases [17]:

The **initialization phase** refers to when the agent accepts a random choice of the hypothesis (such as the index of the element for a dataset, which is an instance number) in the search space. All these 'pointers' were utilized for leading the course of search of the SDS population. Once the phase of initialization is complete, every agent will be given a random hypothesis found in the search space, which is in the **test phase**. Every agent's hypothesis has been individually assessedbased on the objective function. When the hypothesis is evaluated and is successful, the status of the agent is set to active and, if not, inactive. Therefore, every agent will adopt a possible Boolean outcomeat the end of test phase. Lastly, there is the **diffusion phase**; any information on the hypothesis was swapped among the agents. For the purpose of this work, there was a passive recruitment strategy made use of wherein every inactive agent will choose another agent in a random manner. If not, an inactive agent chooses a hypothesis randomly.

There is a new tranquility search algorithm that has come up and is a strong tool to address other complex issues that are similar to the methods of optimization. This is a technique of artificial intelligence that makes use of mechanisms of stochastic computation for determining other near-optimal solutions for one-dimensional and multi-dimensional purposes in accordance with the objective function. This algorithm also functions in a common way and is implemented easily, and is used in an efficient manner in the domains of the class [18]. The advantages of Tranquility Search are: Combining the tranquility search algorithm with that of the other algorithms will be satisfying. This is an excellent type of convergence. It is a great method to get surprising answers. It is suitable for different optimization problems. It is very efficient as an international scheme. It fits in for extensive space. It is robust in dealing with different determinations. It has a high level of efficiency and feasibility to generate global optima. It has a much lower chance of being stuck in the closer optima. The Tranquility search algorithm has a very straightforward concept, and its implementation is connected with the processes of heuristic optimization. It has a reasonable execution time and parameter tuning. It also has scalability, robustness, and adaptability. The disadvantages of the primary tranquility search algorithm were proposed for the multi-objective, single-objective,

and discrete problems. There is not a theoretical converging body. It suffers from the problems of premature convergence. It has the possibility of distribution changes along with all the needs of its generations.

Owing to the fact that the strategy of TM improvisation with the associated parameters of control, PAR, TMCR, and the distance bandwidth are able to determine the algorithm and its performance, it becomes a challenge to choose a suitable strategy for TM improvisation. This should have well-suited parameter values while applying a tranquility search algorithm for solving a problem. The TM refers to the pool of an elite solution that has a key role to play in an algorithm. For the purpose of this work, the SDS has been applied for optimizing the TM in the form of a learning mechanism. This has some agents that are considered and have been explored and exploited using an initialization phase, diffusion phase, and test phase. After this, another new tranquility vector will be generated in the form of an SDS- Tranquility Search Algorithm (SDS-TSA), which is quite effective in identifying better solutions compared to other variants [19].

The steps to be followed for the tranquility-SDS K-Means algorithm are:

Step 1: Initializing the problems along with the algorithm parameters

Step 2: Initializing the tranquility memory by using K-Means clustering

Step 3: Improvising another new tranquility by using the SDS algorithm

Step 4: Updating the tranquility memory

Step 5: Checking the stopping criterion.

3.3 *Proposed Tranquility-Tabu-SDS Algorithm*

The hybrid tranquility algorithm proposed begins with the application of the Tabu Search (TS) as the initial step, after which the tranquillity algorithm is applied to it. In the beginning, the TS begins by the initialization of a TABU list with all different candidate solutions that generate an initial solution compared to the best candidate of the TABU list. In case it is better, it will be added to the TABU list until such time a termination condition is met with. After this, the tranquility search algorithm is applied to initialize the TM that considers the entries of the TABU list in the form of tranquility vectors. Another new and improvised solution is used from the TM. Each component will be based on the TMCR and is taking through by improvising with the TMCR parameter along with a mutation in accordance with the PAR parameter on the completion of each mutation. The best among the quality vectors is chosen at the lowest cost [20].

The Tranquility-Tabu-SDS algorithm will be an extension of the Tranquility-TABU evaluation concept used in a metaheuristic form. At the beginning of this iterative process, the initial population is employed. In this generation, the SDS construction phase is employed. There are different sets of individuals that are called the population-based on which this algorithm will work. This algorithm has been divided into two different subsets spawned with different and distinct mannerisms. A set will be generated by making use of the recommendation operator along with a new selection scheme; it will be a conventional and evolutionary method. The construction phase has been employed as another phase of the SDS. The difference, however, is that there can be a traditional TABU evaluation function that is replaced using a general version of this Tranquility-TABU evaluation function [21].

The steps employed in the Tranquility-TABU-SDS algorithm are as below:

Step 1: Initializing the problem along with the algorithm parameters

Step 2: Initializing the tranquility memory with the TS algorithm

Step 3: Improvising another new tranquility with the SDS algorithm

Step 4: Updating the tranquility memory

Step 5: Checking for the stopping criterion

4. Results and Discussion

The proposed methods SS K-Means, Tranquility K-Means, Tranquility -SS K-Means, and Tranquility-Tabu-SS K-Meansmethods are evaluated using Enron dataset. 90000 emails were used for evaluation of the algorithms.The summary of resultsare tabulated in table 1. The True Positive Rate (TPR) for the known and unknown recipient and False Positive Rate (FPR) for the known and unknown recipient and TP) obtained and the same are shown in figures 2 to 4.

Techniques	SS-K Means	Tranquility K-Means	Tranquility -SS K-Means	Tranquility-Tabu-SS K-Means
TPR for Known recipient	0.8572	0.8547	0.9088	0.9159
TPR for Unknown recipient	0.8718	0.8609	0.8789	0.9086
FPR for Known recipient	0.1282	0.1391	0.1211	0.0914
FPR for Unknown recipient	0.1428	0.1453	0.0912	0.0841

Table 1 Summary of Results

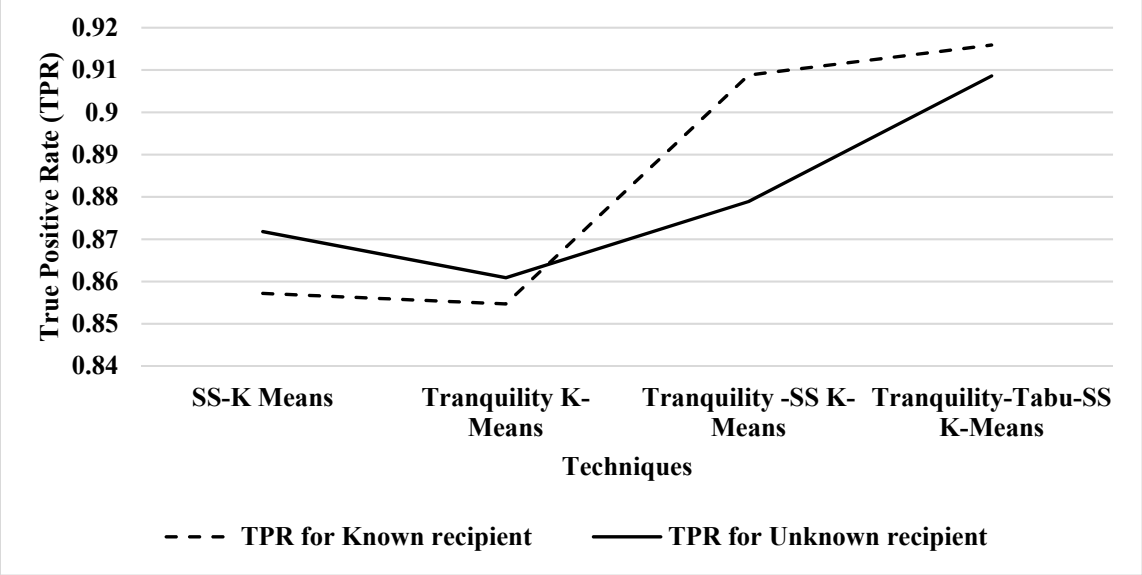

Figure 2 True Positive Rate (TPR) for Tranquility-Tabu-SS K-Means

From figure 2, it is seen that the Tranquility-Tabu-SS K-Means has higher TPR for the known recipient by 6.62% for SS K-Means, by 6.91% for Tranquility K-Means and by 0.78% for Tranquility -SS K-Means, respectively. The Tranquility-Tabu-SS K-Means has higher TPR for the unknown recipient by 4.13% for SS K-Means, by 5.39% for Tranquility K-Means and by 3.32% for Tranquility -SS K-Means, respectively. In Zilberman et al. (2011) [9], the TPR for known recipient is 0.874 whereas for the proposed Tranquility-Tabu-SS K-Means it is 0.9159.

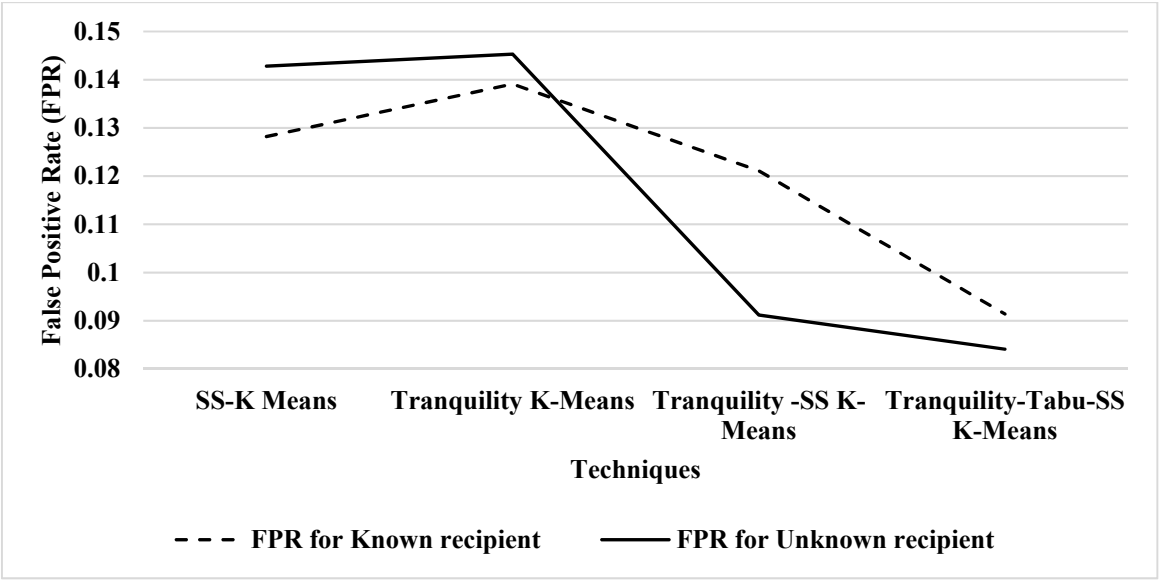

Figure 3 False Positive Rate (FPR) for Tranquility-Tabu-SS K-Means

From figure 3, it is seen that the Tranquility-Tabu-SS K-Means has lower FPR for known recipientsby 33.51% for SS K-Means, by 41.28% for Tranquility K-Means and by 27.95% for Tranquility -SS K-Means, respectively. The Tranquility-Tabu-SS K-Means has lower FPR for the unknown recipient by 51.74% for SS K-Means, by 53.35% for Tranquility K-Means and by 8.1% for Tranquility-SS K-Means, respectively. In Zilberman et al. (2011) [9], the FPR for known recipient is 0.2132 whereas for the proposed Tranquility-Tabu-SS K-Means it is 0.0914.

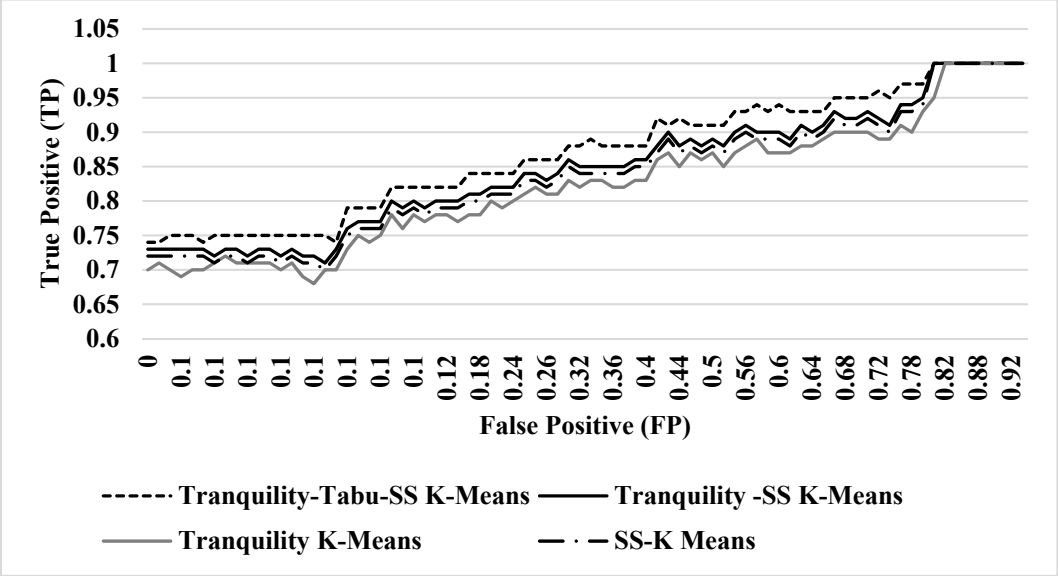

Figure 3 True Positive (TP) Obtained for Tranquility-Tabu-SS K-Means

From figure 3, it is seen that the Tranquility-Tabu-SS K-Means has a higher average TP by 3.66% for SS K-Means, by 5.32% for Tranquility K-Means and by 2.61% for Tranquility -SS K-Means, respectively.

5. Conclusion

DLP refers to a field of research dealing with studying the potential threats to the security of organizational strategies and data that can prevent these threats. Cluster analysis is yet another attractive technique of data mining used in several fields. Another popular algorithm of data clustering is the one called the centre-based clustering algorithm. The K-means is normally used in the form of a clustering method as it is very simple and is of high speed. For the purpose of this work, Tranquility K-Means, Tranquility-TABU-SDS, and Tranquility-SDS K-Means have been proposed. Tranquility search is taken to be the behaviour of the musician

inspired using a soft computing algorithm, as musicians in the process of improvisation attempt at finding the harmony that is the best to their aesthetics, the process of decision variable optimization attempts at being the best vector of the objective function. Another important version of such tranquility search can be the approach that is proposed, the TABU-SDS,which is capable of decreasing the local minima issue and enhancing optimal solutions. The strategy has employed the best neighbours that exist in the current list consisting of the best solutions that are used in the SDS for improvising the algorithm when it is not able to identify a new neighbour and gets stuck in a local minimum. In the phase of construction, the SDS, which is a traditional function of evaluation, is replaced by means of a general version of the function of Tranquility-TABU evaluation. The results demonstrate that the Tranquility-Tabu-SS K-Means has higher TPR for the known recipient by 6.62% for SS K-Means, by 6.91% for Tranquility K-Means and by 0.78% for Tranquility -SS K-Means, respectively. The Tranquility-Tabu-SS K-Means has higher TPR for the unknown recipient by 4.13% for SS K-Means, by 5.39% for Tranquility K-Means and by 3.32% for Tranquility -SS K-Means, respectively.

References

[1] Obar, J. A., & Wildman, S. S. (2015). Social media definition and the governance challenge-an introduction to the special issue. Obar, JA, and Wildman, S.(2015). Social media definition and the governance challenge: An introduction to the special issue. Telecommunications policy, 39(9), 745-750.
[2] Shozi, N. A., & Mtsweni, J. (2017, May). Big data privacy in social media sites. In 2017 IST-Africa Week Conference (IST-Africa) (pp. 1-6). IEEE.
[3] Raman, P., Kayacik, H. G., & Somayaji, A. (2011, June). Understanding data leak prevention. In 6th Annual Symposium on Information Assurance (ASIA'11) (p. 27).
[4] Ojoawo, A. O., Fagbolu, O. O., Olaniyan, A. S., & Sonubi, T. A. (2014). Data leak protection using text mining and social network analysis. Int J Eng Res Dev, 10(12), 14-22.
[5] Gurusamy, V., Kannan, S., & Prabhu, J. R. (2017). Mining the attitude of social network users using k-means clustering. International Journal, 7(5).
[6] Forsati, R., Meybodi, M., Mahdavi, M., & Neiat, A. (2008, December). Hybridization of k-means and harmony search methods for web page clustering. In 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (Vol. 1, pp. 329-335). IEEE.
[7] Mahdavi, M., & Abolhassani, H. (2009). Harmony K-means algorithm for document clustering. Data Mining and Knowledge Discovery, 18(3), 370-391.
[8] Zilberman, P., Dolev, S., Katz, G., Elovici, Y., & Shabtai, A. (2011, July). Analyzing group communication for preventing data leakage via email. In Proceedings of 2011 IEEE international conference on intelligence and security informatics (pp. 37-41). IEEE.
[9] Fan, D., & Wang, J. (2020). An individual-based mean-field model for fake-news spreading on PSO-based networks. International Journal of Modern Physics B, 34(19), 2050172.
[10] Kaur, K., Gupta, I., & Singh, A. K. (2017). A Comparative evaluation of data leakage/loss prevention systems (DLPs). In Proc. 4th Int. Conf. Computer Science & Information Technology (CS & IT-CSCP), Dubai, UAE (pp. 87-95).
[11] Botti-Cebriá, V., del Val, E., & García-Fornes, A. (2020, September). Automatic Detection of Sensitive Information in Educative Social Networks. In Conference on Complex, Intelligent, and Software Intensive Systems (pp. 184-194). Springer, Cham.
[12] Gupta., K and Kush., A (2017) "A Review on Data Leakage Detection for Secure Communication", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-7 Issue-1, October 2017.
[13] Amor, I. B., Benbernou, S., Ouziri, M., Malik, Z., &Medjahed, B. (2016). Discovering best teams for data leak-aware crowdsourcing in social networks. ACM Transactions on the Web (TWEB), 10(1), 1-27.
[14] Katz et al., (2014) Mohassel, P., Rosulek, M., & Trieu, N. (2020). Practical privacy-preserving k-means clustering. Proceedings on Privacy Enhancing Technologies, 2020(4), 414-433.
[15] Mohassel, P., Rosulek, M., & Trieu, N. (2020). Practical privacy-preserving k-means clustering. Proceedings on Privacy Enhancing Technologies, 2020(4), 414-433.
[16] Amiri, B., Hossain, L., &Mosavi, S. E. (2010, October). Application of harmony search algorithm on clustering. In Proceedings of the world congress on engineering and computer science (Vol. 1, pp. 20-22).
[17] Alhakbani, H. A., & Al-Rifaie, M. M. (2016). Exploring Feature-Level Duplications on Imbalanced Data Using Stochastic Diffusion Search. In Multi-Agent Systems and Agreement Technologies (pp. 305-313). Springer, Cham.
[18] Abualigah, L., Diabat, A., &Geem, Z. W. (2020). A Comprehensive Survey of the Harmony Search Algorithm in Clustering Applications. Applied Sciences, 10(11), 3827.
[19] Wu, B., Qian, C., Ni, W., & Fan, S. (2012). Hybrid harmony search and artificial bee colony algorithm for global optimization problems. Computers & Mathematics with Applications, 64(8), 2621-2634.
[20] Alazzam, H., Alhenawi, E., & Al-Sayyed, R. (2019). A hybrid job scheduling algorithm based on Tabu and Harmony search algorithms. The Journal of Supercomputing, 75(12), 7994-8011.
[21] Mandapati, S. (2018). A greedy stochastic diffusion search based fuzzy scheduling in cloud. Journal of Artificial Intelligence Research & Advances, 5(2), 58-68.

Authors Profile

Mr.J.Karthik completed his B.Tech(CSE) from J.N.T.University, Kakinada in 2010 and M.Tech(CSE) from J.N.T.University, Kakinada in 2012. He is currently pursuing Ph.D. in Annamalai University and working as Assistant Professor in Department of Computer Science & Engineering, Gudlavalleru Engineering College, Gudlavalleru..He has published seven research papers in reputed international and two in International conference and it's also available online. His main research work focuses on Data Mining, Information Security, and Cloud Computing. He has 9 years of teaching experience.

Dr.V.Tamizhazhagan working as Assistant Professor in Department of Information Technology, Annamalai University. He has published twelve research papers in reputed International Journals and ten International conferences and it's also available online. He attended & organized various workshops. His main research work focuses on Data Mining, Network Security, Mobile Computing and Computer Networks. He has 16 years of teaching experience. He guided 17 Post graduation, many under graduation projects and guiding 4 Ph.Ds

Dr.Narayana Satyala completed his B.Tech, M.TechandPh.D.fromJ.N.T.University,hyderbad.He is working as Professor in Department of Computer Science Engineering, Gudlavalleru EngineeringCollege,Gudlavalleru..Hehaspublished 13 research papers in reputed international and 2 in International conference and it's also available online. His main research work focuses on Data Mining,Machine Learning. He has 22 years of teaching experience. He guided 12 Post graduation and 20 Under graduation projects.