

A Review of Privacy Preserving in Document Clustering

Miss. Monika Thakor¹ Dharmesh Bhalodiya²

¹P.G. Student ²Assistant Professor

^{1,2}Department of Computer Engineering

^{1,2}Silver-Oak College of Engineering & Technology, Ahmedabad

Abstract— Association Rule Mining from a large amount of data is one of the most important issues in data mining, because the discovered knowledge is commercially valuable. Sometimes companies involved in the similar business are often willing to co-operate each other so that they can perform data mining to extract knowledge from combined datasets. Generally the main objective behind such kind of data mining is mutual gain of all involved parties. But the company dataset contains private or sensitive data. Therefore companies may want certain strategic or private data called sensitive patterns not to be published in the database. Therefore, before the database is released for sharing, some sensitive patterns have to be hidden in the database because of privacy or security concerns. To solve this problem, sensitive-knowledge-hiding (association rules hiding) problem has been discussed in the research community working on security and knowledge discovery. A lot of research has been completed to solve the problem. In this thesis, we will introduce an efficient algorithm to protect sensitive information.

Key words: Document Clustering, Clustering Method, Hierarchical Clustering

I. INTRODUCTION

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. In other words, the goal of a good document clustering scheme is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances (using an appropriate distance measure between documents). A distance measure (or, dually, similarity measure) thus lies at the heart of document clustering. Clustering is the most common form of unsupervised learning and this is the major difference between clustering and classification. No super-vision means that there is no human expert who has assigned documents to classes. In clustering, it is the distribution and makeup of the data that will determine cluster membership. Clustering is sometimes erroneously referred to as automatic classification; however, this is inaccurate, since the clusters found are not known prior to processing whereas in case of classification the classes are pre-defined.

A. Document Clustering:

Document clustering is a more specific technique for unsupervised document organization, it is generally considered to be a centralized process. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories

Document clustering is considered as a centralized process has been in use in a number of different areas of text mining and information retrieval. Clustering can be considered the most important unsupervised learning problem so, as every other problem of this kind; it deals

with finding a structure in a collection of unlabeled data. A cluster is therefore a collection of objects which are “similar” between them and are Document clustering is considered as a centralized process has been in use in a number of different areas of text mining and information retrieval. Clustering can be considered the most important unsupervised learning problem so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A cluster is therefore a collection of objects which are “similar” between them.

- scalability;
- dealing with different types of attributes;
- discovering clusters with arbitrary shape;
- minimal requirements for domain knowledge to determine input parameters;
- ability to deal with noise and outliers;
- insensitivity to order of input records;
- high dimensionality;
- interpretability and usability

II. CLUSTERING METHODS

A. K-Means:

K-means is the most important flat clustering algorithm. The objective function of Kmeans is to minimize the average squared distance of objects from their cluster centers, where a cluster center is defined as the mean or centroid μ of the objects in a cluster C:

The ideal cluster in K-means is a sphere with the centroid as its center of gravity. Ideally, the clusters should not overlap. A measure of how well the centroids represent the members of their clusters is the Residual Sum of Squares (RSS), the squared distance of each vector from its centroid summed over all vectors

K-means can start with selecting as initial clusters centers K randomly chosen objects, namely the seeds. It then moves the cluster centers around in space in order to minimize RSS. This is done iteratively by repeating two steps until a stopping criterion is met

- 1) Reassigning objects to the cluster with closest centroid
- 2) Recomputing each centroid based on the current members of its cluster.

We can use one of the following termination conditions as stopping criterion

- A fixed number of iterations I has been completed.
- Centroids μ do not change between iterations.
- Terminate when RSS falls below a pre-established threshold.

B. Hierarchical Clustering:

Hierarchical clustering approaches attempt to create a hierarchical decomposition of the given document collection thus achieving a hierarchical structure. Hierarchical methods

are usually classified into Agglomerative and Divisive methods depending on how the hierarchy is constructed.

III. METHODOLOGY

A. Agglomerative Methods:

Agglomerative methods start with an initial clustering of the term space, where all documents are considered representing a separate cluster. The closest clusters using a given inter-cluster similarity measure are then merged continuously until only 1 cluster or a predefined number of clusters remain. Simple Agglomerative Clustering Algorithm:

- Compute the similarity between all pairs of clusters i.e. calculate a similarity matrix whose ij entry gives the similarity between the I and j clusters.
- Merge the most similar (closest) two clusters.
- Update the similarity matrix to reflect the pairwise similarity between the new cluster and the original clusters.
- Repeat steps 2 and 3 until only a single cluster remains.

B. Divisive Clustering:

Divisive clustering algorithms start with a single cluster containing all documents. It then continuously divides clusters until all documents are contained in their own cluster or a predefined number of clusters are found. Agglomerative algorithms are usually classified according to the inter-cluster similarity measure they use. The most popular of these are single-link, complete-link and group average. In the *single link* method, the distance between clusters is the minimum distance between any pair of elements drawn from these clusters (one from each), in the *complete link* it is the maximum distance and in the *average link* it is correspondingly an average distance

C. Hierarchical Analysis Model:

A hierarchical clustering algorithm creates a hierarchical decomposition of the given set of data objects. Depending on the decomposition approach, hierarchical algorithms are classified as agglomerative (merging) or divisive (splitting). The agglomerative approach starts with each data point in a separate cluster or with a certain large number of clusters. Each step of this approach merges the two clusters that are the most similar. Thus after each step, the total number of clusters decreases. This is repeated until the desired number of clusters is obtained or only one cluster remains. By contrast, the divisive approach starts with all data objects in the same cluster. In each step, one cluster is split into smaller clusters, until a termination condition holds. Agglomerative algorithms are more widely used in practice. Thus the similarities between clusters are more researched.

D. Privacy Preserving:

Privacy preserving data mining is a new investigation in data mining and statistical databases. In PPDM data mining algorithms are analyzed for side effects obtain in data privacy. Two fold consideration in privacy preserving data mining. First is sensitive raw data that are kept secure from unauthorized access like identifiers, names ,addresses should be modified from original database in order for receiver of data not to be able to compromise another person's privacy. Second is sensitive knowledge is excluded

that can be mined from a database by using data mining algorithms as such type of knowledge compromises data privacy.

IV. LITERATURE REVIEW

Document clustering is widely applicable in areas such as search engines, web mining, information retrieval, and topological analysis. Most document clustering methods perform several preprocessing steps including stop words removal and stemming on the document set. Each document is represented by a vector of frequencies of remaining terms within the document. Some document clustering algorithms employ an extra preprocessing step that divides the actual term frequency by the overall frequency of the term in the entire document set.

[1] Tiancheng Li, Ninghui Li- In this paper, They present a novel technique called slicing, which partitions the data both horizontally and vertically. They show that slicing preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of slicing is that it can handle high-dimensional data. They show how slicing can be used for attribute disclosure protection and develop an efficient algorithm for computing the sliced data that obey the ϵ -diversity requirement. Workload experiments confirm that slicing preserves better utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute. Experiments also demonstrate that slicing can be used to prevent membership disclosure.

This paper presents a new approach called slicing to privacy preserving micro data publishing. Slicing overcomes the limitations of generalization and bucketization and preserves better utility while protecting against privacy threats. They illustrate how to use slicing to prevent attribute disclosure and membership disclosure. Experiments show that slicing preserves better data utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute. The general methodology proposed by this work is that: before anonymizing the data, one can analyze the data characteristics and use these characteristics in data anonymization.. The rationale is that one can design better data anonymization techniques when we know the data better.

[2] TamirTassa- One research problem that this study suggests was described in Section 3; namely, to devise an efficient protocol for inequality verifications that uses the existence of a semi-honest third party. Such a protocol might enable to further improve upon the communication and computational costs of the second and third stages of the protocol of as described in Sections 3 and 4. Other research problems that this study suggests is the implementation of the techniques presented here to the problem of distributed association rule mining in the vertical setting the problem of mining generalized association rules and the problem of subgroup discovery in horizontally partitioned data. The protocol that we propose here computes a parameterized family of functions, which we call threshold functions, in which the two extreme cases correspond to the problems of computing the union and intersection of private subsets.

Those are in fact general-purpose protocols that can be used in other contexts as well. Another problem of secure multiparty computation that we solve here as part of our

discussion is the set inclusion problem; namely, the problem where Alice holds a private subset of some ground set, and Bob holds an element in the ground set, and they wish to determine whether Bob's element is within Alice's subset, without revealing to either of them information about the other party's input beyond the above described inclusion.

[3] Shikha Sharma- In this thesis they remove this drawback. In other techniques non-sensitive rules hidden (falsely) as a side effect and artificial rules falsely generated in other rule hiding techniques. In this paper we present a new approach that necessarily changes few transactions in the transaction database by decreasing support or confidence of sensitive rules without any side effect. An Apriori algorithm is one of the known techniques that is used. In this research, this technique is used for privacy data identification and extraction from printed documents. In this research they point the problem of discovering association rules for various privacy types from printed documents. An association rule expresses the dependence of a set of attribute-value pairs and upon another set of items (item set). The mining of association rules is performed in two stages: The frequent sets of items from the data discovery and association rules generation from the frequent item sets. Searching of these frequent item sets is in general combinatorial expensive task. Association rule mining has a broad range of applicability. It was first introduced to find the association between items in supermarket transactions for promotion of sales, arrangement of associated items accordingly, to increase profits etc. From experimental results we see that our approach is better in the way that it hides any rule which cannot be hidden by some of the previous works. We see in the example that proposed method is hiding the given association rules TSHIRT->JEANS (with sensitive items on the left hand side of the rule) without any side effect. The aim of this research release non-sensitive item sets while keeping sensitive items private.

[4] QusayBsoul, JuhanaSalim- This paper detects and identifies some limitations in Crime Document Clustering. First off, it addresses the fault detection and identification in the k-means algorithm, then, it examines the weakness of extraction terms from documents as it is stated above. Therefore, this study aims to enhance the reliability of Document Clustering of crime document by efficient k-means as well as the extraction features of crime document. Furthermore, it is used for crime document clustering, and its results are the best testimony for its efficiency as it aims to enhance the kmeans algorithm for Document Clustering as well as the extracting of information which group topics/events of crimes can outperform the original Document Clustering and other Document Clustering based on two criteria of time and performance. Besides, we look forward that our suggestion of crime document clustering enhances the performance and the effectiveness.

[5] Himanshu Gupta- The method represented in this paper computes the value of k automatically and with great precision. By using refinement by feature voting increase the efficiency greatly. Method is both efficient and accurate and creates good quality clusters.

[6] HemalathaImmandhi-The work in this paper is motivated by investigations from the above and similar

research findings. It appears to us that the nature of similarity measure plays a very important role in the success or failure of a clustering method. Our first objective is to derive a novel method for measuring similarity between data objects in sparse and high dimensional domain, particularly text documents. From the proposed similarity measure, we then formulate new clustering criterion functions and introduce their respective clustering algorithms, which are fast and scalable like kmeans, but are also capable of providing high-quality and consistent performance

This report has a proposal of a new hierarchical clustering algorithm based on the overlap rate for cluster merging. The experience in general data sets and a document set indicates that the new method can decrease the time cost, reduce the space complexity and improve the accuracy of clustering. Specially, in the document clustering, the newly proposed algorithm measuring result show great advantages. The hierarchical document clustering algorithm provides a natural way of distinguishing clusters and implementing the basic requirement of clustering as high within-cluster similarity and between-cluster dissimilarity.

[7] Todsanaichumwatana-As can be observed from the results, the accuracy of the proposed method is up to 83.25%, meanwhile using the SOM based documents clustering using single words and the hierarchical clustering approach provide the accuracies 72.21% and 79.75% respectively. The hierarchical clustering approach also created many small clusters that containing only a few documents. As a result, an improvement was demonstrated using frequent max sub-strings rather than single words as features. This proposed technique also does not require any pre-processing technique to extract the frequent max substrings. Meanwhile, the SOM based documents clustering using single words and the hierarchical based document clustering using single words require word segmentation to extract the single words.

This paper describes a non-segmented document clustering method using self-organizing map (SOM) and frequent max substring technique to improve the efficiency of information retrieval. We first use the frequent max substring technique to discover patterns of interest, called frequent max substrings, rather than individual words from Thai text documents, and these frequent max sub-strings are then used as indexing terms with their number of occurrences to form a document vector. SOM is then applied to generate the document cluster map by using the document vector. The experiment studies and comparison results on clustering the 50 Thai text documents is presented in this paper.

[8] Bashar Aubaidan-The aim of this study is to conduct a comparative study of two main clustering algorithms, namely k-means and k-means++. The method of this study includes a preprocessing phase, which in turn involves tokenization, stop-words removal and stemming. In addition, we evaluate the impact of the two similarity/distance measures (Cosine similarity and Jacquard coefficient) on the results of the two clustering algorithms. Experimental results on several settings of the crime data set showed that by identifying the best seed for initial cluster centers, k-mean++ can significantly (with the significance interval at 95%) work better than k-means. These results

demonstrate the accuracy of k-mean++ clustering algorithm in clustering crime documents.

This study was aimed to investigate the best similarity in k-means and k-means++ for crime document and to evaluate and compare the performance of k-means and k-means++ in clustering. In this study, we had used crime Dataset collected from Bernama news and have tested six categories of topics. Based on the results in section 4, the K-means++ algorithm has the best results with Cosine similarity compared to Jaccard similarity. Experimental method, based on K-means ++, has been proved to be better and more accurate than the k-means clustering, in crime document clustering the results show that the k-means++ outperforms the k-means and that cosine similarity performs better than the Jaccard coefficient. The reason for this is due to the fact that the k-means identifies the first initial centroid randomly, while the k-means++ algorithm selects the second initial centroid mathematically through probability proportional to the square of the distance over summation of the square distance for the current point. As for the performance of the cosine similarity, it outperformed the Jaccard coefficient because it is independent of document length and the data set consisted of documents with different lengths.

[9] TanushriPotphode- By doing this literature survey we studied that the existing system have some problem such as accuracy, required more time for finding relevant document from huge amount of clusters that's why to overcome this problem we proposed new text clustering algorithm such as K-representative algorithm which will give us the better computer forensic analysis. The main idea of K-representative algorithm is to use the relative attribute frequencies of the clusters mode in the dissimilarity measures in the K-mode objective function. It has been shown that K-representative algorithm is very efficient. Due to the modification proposed in forming representatives for clusters of categorical objects, the dissimilarity between a categorical object and the representative of a cluster

In this paper our survey shows how different document clustering techniques are used for digital forensic analysis with different phases involve in it. In addition to this, we present an approach for implementation of enhance text clustering algorithm which will form clusters on the basis of relative match. It gives better results and improves the accuracy of clustering technique. By using this approach searching time for finding relevant document from huge amount of datasets will be reduced and improve the efficiency of forensic analysis.

[10]RakshaK.Mundhe-In proposed approach the forensic analysis is done very systematically i.e. retrieved data is in unstructured format get particular structure by using high quality well known algorithm and automatic cluster labeling method. Two relative validity indexes were used to automatically estimate the number of clusters with automatic labeling to it; which makes it very easy to retrieve most relevant information for forensic analysis.

In this paper the high quality clustering algorithm, and methods used which will provide the automatic labeling to the cluster, and will provide the indexing to text, doc, and pdf file.

V. CONCLUSION

we present a novel approach that modifies the database to hide sensitive rules with limited side effects. Proposed method classifies all the valid modifications such that every class of modifications is related with the sensitive rules, non-sensitive rules that can be affected after the modifications. It modifies the transactions in an order so that both the numbers of hidden sensitive rules and modified entries are considered. In most cases, all the sensitive rules are hidden without false rules generated or lost rule.

REFERENCE

- [1] A New Approach for Privacy Preserving Data Publishing Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jian Zhang, Member, IEEE, and Ian Molloy, MARCH 2012
- [2] Secure Mining of Association Rules in Horizontally Distributed Databases TamirTassa, APRIL 2014
- [3] An Extended Method for Privacy Preserving Association Rule Mining Shikha Sharma, International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 10, October 2012
- [4] An Intelligent Document Clustering Approach to Detect Crime Patterns QusayBsoul, JuhanaSalim, LailatulQadriZakaria, ICEEI 2013
- [5] k-means Based Document Clustering with Automatic "k" Selection and Cluster Refinement Himanshu Gupta *et al*, International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 5, May- 2014
- [6] Evaluation of Similarities Measure in Document Clustering HemalathaImmandhi, International Journal of Science and Research (IJSR) January 2014
- [7] A SOM-Based Document Clustering Using Frequent Max Substrings for Non-Segmented Text Todsanaichumwatana, KokWai Wong, Hong Xie, Received March 25th, 2010; revised July 15th, 2010; accepted July 30th, 2010.
- [8] COMPARATIVE STUDY OF K-MEANS AND K-MEANS++ CLUSTERING ALGORITHMS ON CRIME DOMAIN Bashra Aubaidan, MasnizahMohd and Mohammed Albared, Journal of Computer Science 10 (7): 1197-1206, 2014 ISSN: 1549-3636 © 2014 Science Publications
- [9] An Empirical Approach for Document Clustering in Forensic Analysis: A Review TanushriPotphode, Prof. AmitPimpalkar, International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 11, November 2014
- [10] Information Retrieval Using Document Clustering for Forensic Analysis RakshaK.Mundhe, AnkushMaind, R.B.Talmale, ISSN (Online): 2347 - 2812, Volume-2, Issue -5, 2014