A Comprehensive Analysis on Text Mining Using Hierarchical Clustering Algorithm

Shravan Kumar[1], Jawahar Kumar[2], Raghvendra Kumar[3]

[1, 2, 3] M. Tech Computer Science & Engineering, Lakshmi Narain College of Technology
Jabalpur (M.P), INDIA,
shravan51090[at]gmail.com
jawahar11088[at]gmail.com
Raghvendraagrawal7[at]gmail.com

Abstract: Data mining is also known as knowledge discovery process, is the analysis process of text, word & data through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining technique predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining technique can answer business questions that traditionally were too time consuming to resolve. It help to search scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data mining technique derives its name from the similarities between searching for important information in a large complex database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides. Data mining is a powerful new technology with great potential to help companies focus on the most important information in the data they have collected about the behavior of their customers and potential customers. It discovers information within the data that queries and reports can't effectively reveal.

Keywords: Data mining**,** Text mining, clustering algorithm.

1. Introduction

Today we live in a internet world and have a large amount of data wand they are stored in a data base and in a database due to large data it is very critical job to search a particular data. After all we take the help of data mining technique and it give a particular information about search which we want to search. Text mining is a process which support to burgeoning new field that are used to search meaningful information from natural language text. It may be loosely characterized as the process of analyzing text to extract information that is useful for particular purposes. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with algorithmically. Today, text is the most common thing for the formal share of information. The area of text mining usually deals with texts which are required to user whose function is the communication of factual information or opinions, and the motivation for trying to extract information from such text automatically is compelling—even if success is only partial. The phrase "text mining" is generally used to search a word and any system that analyzes large and complex quantities of natural language text and detects lexical patterns in an attempt to extract probably useful (although only probably correct) information. The discovery of multi-dimensional data has proceeded at a likely to explode rate in many disciplines with the advance of recent technology, which greatly increases the challenges of comprehend and interpreting the resulting mass of data. Existing data analysis techniques have complexity in handling multi-dimensional data. Multi-dimensional data has been a challenge for data analysis because of the inherent in sufficiency of the points. The architecture of a data mining technique have the following major components ,database, data warehouse, or other in sequence repository; their server which is responsible for fetching the applicable data based on the user's data mining request; knowledge base data which is used to channel the search, or evaluate the interestingness of resultant patterns; data mining engine which consists of a large set of functional modules for tasks; pattern assessment module which interacts with the data mining modules so as to focus the search towards motivating patterns; and graphical user interface which communicates between users and the data mining system, allowing the user interaction with system. According to the rapid growth of large digital data made available in recent years, knowledge discovery and data mining have attracted a great deal of attention with an imminent need for turning such data into useful information and knowledge. Many process & applications, such as data analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. Knowledge discovery can be viewed as the process of nontrivial extraction of information from large databases, information that is implicitly presented in the data, previously unknown and potentially useful for users. Data mining system is therefore an essential step in the process of knowledge discovery in databases.

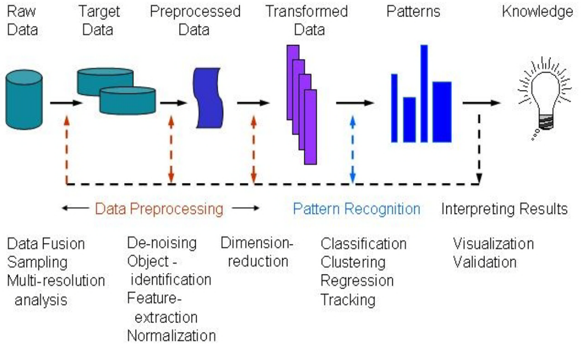

Fig. 1.1 process of data miming

2. Introduction on Cluster Analysis

Clustering and classification are both fundamental tasks in Data Mining. Classification is used mostly as a supervised learning method, clustering for unsupervised learning (some clustering models are for both). The goal of clustering is descriptive, that of classification is predictive. Since the goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is intrinsic. In classification tasks, however, an important part of the assessment is extrinsic, since the groups must reflect some reference set of classes. Clustering groups data instances into subsets in such a manner that similar
instances are grouped together, while different instances belong to different groups. The instances are thereby organized into an efficient representation that characterizes the population being sampled. Formally, the clustering structure is represented as a set of subsets
$C = C1; : : : ;Ck$ of S, such that:
$S = Sk$
i=1 Ci and $Ci \setminus Cj$ =; for i 6= j. Consequently, any instance in S belongs to exactly one and only one subset. Clustering of objects is as ancient as the human need for describing the salient characteristics of men and objects and identifying them with a type. Therefore, it embraces various scientific disciplines: from mathematics and statistics to biology and genetics, each of which uses different terms to describe the topologies formed using this analysis.

Cluster analysis is the process of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

3. Issued Algorithms

3.1 Partitioning Algorithms

The PAM algorithm was developed by Leonard Kaufman and Peter J. Rousseeuw, and this algorithm is very similar to K-means, mostly because both are partitional algorithms, in other words, both break the dataset into groups (clusters), and both work by trying to minimize the error, but PAM works with Medoids, that are an entity of the dataset that represent the group in which it is inserted, and K-means works with Centroids, that are artificially created entity that represent its cluster.

The PAM algorithm partitions the dataset of n objects into k clusters, where both the dataset and the number k is an input of the algorithm. This algorithm works with a matrix of dissimilarity, whose goal is to minimize the overall dissimilarity between the representants of each cluster and its members. The algorithm uses the following model to solve

the problem:

Subject to:

1. $\Sigma_{i=1}^{n} z_{ij} = 1$, j = 1,2,...,n
2. $z_{ij} \leq y_i$, i, j = 1,2,...,n
3. $\Sigma_{i=1}^{n} y_i = k$, k = number of clusters
4. y_i , z_{ij} € {0,1} , i, j = 1,2,...,n

3.2 Hierarchical Algorithms

A hierarchical clustering method consists of grouping data objects into a tree of clusters. There are two main types of techniques: a bottom-up and a top-down approach. The first one starts with small clusters composed by a single object and, at each step, merge the current clusters into greater ones, successively, until reach a cluster composed by all data objects. The second approach use the same logic, but to the opposite direction, starting with the greatest cluster, composed by all objects, and split it successively into smaller clusters until reach the singleton groups. Besides the strategies, other important issue is the metrics used to build (merge or split) clusters. Such metrics can be different distance measures, described next section.

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy. There are two types of hierarchical clustering, Divisive and Agglomerative.
In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram. In the general case, the complexity of agglomerative clustering is , which makes them too slow for large data sets. Divisive clustering with an exhaustive search is , which is even worse. However, for some special cases, optimal efficient agglomerative methods (of complexity) are known: SLINK for single-linkage and CLINK for complete-linkage clustering

3.3 Agglomerative algorithms

Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc. The classic example of this is species taxonomy. Gene expression data might also exhibit this hierarchical quality (e.g. neurotransmitter gene families). Agglomerative hierarchical clustering starts with every single object (gene or sample) in a single cluster. Then, in each successive iteration, it agglomerates (merges) the closest pair of clusters by satisfying some similarity criteria, until all of the data is in one cluster.

The hierarchy within the final cluster has the following properties:

- Clusters generated in early stages are nested in those generated in later stages.
- Clusters with different sizes in the tree can be valuable for discovery.

A Hybrid Hierarchical Clustering is a clustering technique that trys to combine the best characteristics of both types of Hierarchical Techniques (Agglomerative and Divisive). Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering is therefore called hierarchical agglomerative clustering or HAC . Top-down clustering requires a method for splitting a cluster. It proceeds by splitting clusters recursively until individual documents are reached.

3.4 Divisive Hierarchical Algorithms

This variant of hierarchical clustering is called top-down clustering or divisive clustering. We start at the top with all documents in one cluster. The cluster is split using a flat clustering algorithm. This procedure is applied recursively until each document is in its own singleton cluster.

Top-down clustering is conceptually more complex than bottom-up clustering since we need a second, flat clustering algorithm as a ``subroutine''. It has the advantage of being more efficient if we do not generate a complete hierarchy all the way down to individual document leaves. For a fixed number of top levels, using an efficient flat algorithm like K-means, top-down algorithms are linear in the number of documents and clusters. So they run much faster than HAC algorithms, which are at least

4. Conclusions

The fast-growing, tremendous quantity of data has far exceeded our human ability for comprehension exclusive of powerful tools. It is really important to design the tools to extract the valuable knowledge entrenched in the vast amounts of data. This dissertation focuses on successful and efficient mining of novel, interesting and important patterns from real data sets. To be specific, we study the following four problems:
1. Shrinking-based data pre-processing approach and shrinking-based clustering algorithm.
2. Shrinking-based dimension reduction approach
3. Iteratively detecting clusters and outliers based on their Inter-relationship, and on the intra-relationship within clusters, and within outliers, respectively
4. Indexing time-related multi-dimensional data sets. We have calculated four interesting problems in clustering, outlier detection and indexing multi-dimensional data.

In the future, we will extend our work along the following directions:
1. Cluster and outlier detection design in subspace;
2. Dynamically insertion for indexing structure of time related multi-dimensional data
3. Combination of fuzzy clustering and shrinking-based data analysis approaches.

References

[1] White D.A., Jain R. Similarity Indexing with the SS-tree. In

[2] C.C. Aggarwal, P. Yu. Finding generalized projected clusters in high dimensional spaces. In Proceedings of the ACM SIGMOD CONFERENCE on Management of Data, pages 70–81, Dallas, Texas, 2000.

[3] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. Lecture Notes in Computer Science, 1973, 2001.

[4] Charu C. Aggarwal, C. Procopiuc, J.L. Wolf, P. Yu, and J.S. Park. Fast algorithms for projected clustering. In Proceedings of the ACM SIGMOD CONFERENCE on Management of Data, pages 61–72, Philadelphia, PA, 1999.

[5] Charu C. Aggarwal, Philip S. Yu. Outlier detection for high dimensional data. In SIGMOD Conference, 2001.

[6] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In Proceedings of the ACM SIGMOD Conference on Management of Data, page 94–105, Seattle, WA, 1998.

[7] N. Ahuja. Dot pattern processing using voronoi neighborhoods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 4(3):336–343, May 1982.

[8] Ankerst M., Breunig M. M., Kriegel H.-P., Sander J. OPTICS: Ordering Points To Identify the Clustering Structure. Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'99), Philadelphia, PA, pages 49–60, 1999.

[9] Bar-Joseph, Z., Demaine, E.D., Gifford, D.K., Srebro, N., Hamel, A.M. and Jaakkola, T.S. K-ary clustering with optimal leaf ordering for gene expression data. Bioinformatics,19(9):1070–1078, 2003.

[10] S. D. Bay. The UCI KDD Archive [http://kdd.ics.uci.edu].University of California, Irvine, Department of Information and Computer Science.

[11] Beckmann N. and Kriegel H.P. and Schneider R. and Seeger B. The R*-tree: an Efficient and Robust Access Method for Points and Rectangles. In Proceedings of ACM-SIGMOD International Conference on Management of Data, pages 322–331, Atlantic City, NJ, May 1990.

[12] Ben-Dor, A., Shamir, R., Yakhini, Z. Clustering gene expression patterns. Journal of Computational Biology, 6(3/4):281–297, 1999.

[13] Bohm C. Berchtold S., Kriegel H. The Pyramid-Technique: Towards Breaking the Curse of Dimensionality. In Proceedings of the 1998 ACM SIGMOD International.

[14] Conference on Management of Data, pages 142–153, Seattle, Washington, 1998.

[15] P. Berkhin. Survey of clustering data mining techniques.Technical report, Accrue Software, San Jose, CA, 2002.

[16] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is

[17] "nearest neighbor" meaningful? In International Conference on Database Theory 99, pages 217–235,