MJRET
Open Access

Multidisciplinary Journal of Research in Engineering and Technology

Journal homepage: www.mjret.in

ISSN:2348 - 6953

DATA CLUSTERING USING CLUSTER PATTERN ANALYSIS

S.J.Jogdand, P.S.Badhe, S.S.Deshmukh, D.S.Shambale, V.P.Rathod, Prof.R.R.Gavli

Computer Department, Bhivrabai Sawant Polythecnic
Pune, India

Abstract- *In today's world, we need to analyse and extract information from the data. The grouping is one of those analyses Method that consists in the distribution of data in groups of identical objects each group is known as a group, which consists of objects that have affinity within the cluster disparity with objects in other groups. This paper is intended examine and evaluate different data grouping algorithms. The Two main categories of cluster approaches are partition and hierarchical grouping. The algorithms discussed here are: k-means clustering algorithm, hierarchical clustering algorithm, density-based clustering algorithm, self-organized map Algorithm and grouping algorithm for maximizing expectations. All the mentioned algorithms are explained and analysed based on in factors such as the size of the data set, the type of data set, Number of created clusters, quality, accuracy and performance. This paper also provides information on the tools that they are used to implement cluster approaches.*

Keywords*- Clustering, K-means clustering algorithm, Pattern analysis.*

1. INTRODUCTION

A large amount of data is collected in different databases due to advanced data collection methods. The request to group important data and extract useful information from data increases. Clustering is the distribution of data in groups of identical objects that have affinity within the cluster and disparity with objects in other groups. The characterization of data in a smaller number of groups will surely lead to a loss in some details, but the data will be interpreted. Represents data objects for a smaller number of cluster numbers and, therefore, model data using their own clusters. The conglomerate analysis is the arrangement of a set of models (usually shown as a measurement vector or a point in a multidimensional space) in clusters based on similarity. Models within the same group are closely related to data in adjacent groups. Here, it is necessary to know the difference between the unsupervised classification and the supervised classification that lies between grouping and discriminatory analysis. In the supervised approach, we are given a series of pre-classified elements; the fact is to label an article just compared, but without labeling. The elements that are already labeled are provided to know the description of the classes that will help us to label a new article. In an unsupervised approach, we will receive a collection of unlabeled articles to classify them into valid groups. The application of grouping approaches has increased considerably in the areas of Artificial Intelligence, pattern recognition, image processing, medicine, marketing, data extraction, image or data compression, statistics, etc. Some of the researchers have improved existing data clustering algorithms, some of the others have designed new methods for grouping data, and some academics have examined and analyzed different methods of data clustering. The goal of clustering is to determine the intrinsic grouping in an untagged dataset. In this document, the software used for the implementation of clustering approaches is analyzed. The dataset that can be used for analysis has also been elaborated in this document. Different data grouping algorithms are explained, analyzed according to the parameters that have been studied up to now.

2. MOTIVATION

In introductory part for the study of Text clustering, their application, which algorithm used for that and the different types of model, I decided to work on the Text clustering which is used for data analysis lot of work done on that application and that the technique used for that application is Text clustering using traditional way. Approaches to the state of the art to classify data it can be used to identify subset of data instances. However, they suffer from low accuracy.

3. RELATED WORK

Literature survey is the most important step in any kind of research. Before start developing we need to study the previous papers of our domain which we are working and on the basis of study we can predict or generate the drawback and start working with the reference of previous papers."In this section, we briefly review the related work on Clustering system and their different techniques.

J.-T. Chien, describe the “Hierarchical theme and topic modeling,” in that Taking into account hierarchical data sets in the body of text, such as words, phrases and documents, we perform structural learning and we deduce latent themes and themes for sentences and words from a collection of documents, respectively. The relationship between arguments and arguments in different data groupings is explored through an unsupervised procedure without limiting the number of clusters. A tree branching process is presented to draw the proportions of the topic for different phrases. They build a hierarchical theme and a thematic model, which flexibly represents heterogeneous documents using non-parametric Bayesian parameters. The thematic phrases and the thematic words are extracted. In the experiments, the proposed method is evaluated as effective for the construction of a semantic tree structure for the corresponding sentences and words. The superiority of the use of the tree model for the selection of expressive phrases for the summary of documents is illustrated [1].

Bernardini, C. Carpineto, and M. D’Amico, describe the “Full-subtopic retrieval with keyphrase-based search results clustering,” in that Consider the problem of restoring multiple documents that are relevant to the individual sub-topics of a given Web query, called "full child retrieval". To solve this problem, they present a new algorithm for grouping search results that generates clusters labelled with key phrases. The key phrases are extracted generalized suffix tree created by the search results and merge through a hierarchical agglomeration procedure improved grouping. They also introduce a new measure to evaluate the performance of full recovery sub-themes, namely "look for secondary arguments length under the sufficiency of k documents". they have used a test collection specifically designed to evaluate the recovery of the sub-themes, they have found that our algorithm has passed both other clustering algorithms of existing research results as a method of redirecting search results underline the diversity of results (at least for k> 1, that is when they are interested in recovering more than one relevant document by sub-theme) [2].

T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, describe the “Self-organization of a massive document collection,” this paper describes the implementation of a system that can organize large collections of documents based on textual similarities. It is based upon the self-organized map (SOM) algorithm. Like the feature vectors for documents, the factual portrayals of their vocabularies are utilized. The main objective of our work was to resize the SOM algorithm in order to handle large amounts of high-dimensional data. In a practical experiment, they mapped 6 840 568 patent abstracts in a SOM of 1.002.240 nodes. As characteristic vectors, we use vectors of 500 stochastic figures obtained as random projections of histograms of weighted words [3].

K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram, describe the “A hierarchical monothetic document clustering algorithm for summarization and browsing search results,” in that Organizing Web search results in a hierarchy of topics and secondary topics makes it easy to explore the collection and position the results of interest. In this paper, they propose a new hierarchical monarchic grouping algorithm to construct a hierarchy of topics for a collection of search results retrieved in response to a query. At all levels of the hierarchy, the new algorithm progressively identifies problems in order to maximize coverage and maintain the distinctiveness of the topics. They refer to the algorithm proposed as DisCover. The evaluation of the quality of a hierarchy of subjects is not a trivial task, the last test is the user's judgment. They have used various objective measures, such as coverage and application time for an empirical comparison of the proposed algorithm with two other monotetic grouping algorithms to demonstrate its superiority. Although our algorithm is a bit more computationally than one of the algorithms, it generates better hierarchies. Our user studies also show that the proposed algorithm is superior to other algorithms as a tool for summary and navigation [4].

R. Xu and D. Wunsch, describe the “Survey of clustering algorithms,” in that Data analysis plays an indispensable role in understanding the various phenomena. Conglomerate analysis, primitive exploration with little or no previous knowledge, consists of research developed in a wide variety of communities. Diversity, on the one hand, provides us with many tools. On the other hand, the profusion of options causes confusion. They have examined the grouping algorithms for the data sets that appear in statistics, computer science and machine learning and they illustrate their applications in some reference datasets, the problem of street vendors and bioinformatics, and a new field that attracts intense efforts. Various closely related topics, proximity measurement and cluster validation are also discussed [5].

4. EXISTING APPROACH

A lot of work has been done in this field thanks to its extensive use and applications. This section mentions some of the approaches that have been implemented to achieve the same purpose. These works are mainly differentiated from the algorithm for clustering systems.

In another research, to access the relevant information from mass of data is very difficult and time consuming task as every day mass of information increases because of digital world.

Every day, the mass of information available to us increases. This information would be irrelevant if our ability to efficiently access did not increase as well. Automated text clustering provide us with maximum benefit that allow us to search, sort, index, store, and analyze the available data. It also allows us to find in desired information in a reasonable time.

As my point of view when I studied the papers the issues are related to clustering techniques.

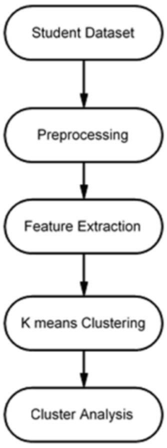
Disadvantages:

- Not worked on automated text classification.
- Everything done manually
- Time consuming process
- No repository Available
- Domain wise analysis is time consuming

5. PROPOSED APPROACH

In another research, to access the relevant information from mass of data is very difficult and time consuming task as every day mass of information increases because of digital world. Every day, the mass of information available to us increases. This information would be irrelevant if our ability to efficiently access did not increase as well. Automated text clustering provide us with maximum benefit that allow us to search, sort, index, store, and analyze the available data. It also allows us to find in desired information in a reasonable time. As my point of view when I studied the papers the issues are related to clustering techniques.

6. PROPOSED SYSTEM DIAGRAM



7. CONCLUSION

As an important tool for data exploration, cluster analysis examine unlabeled data, both by building a hierarchy structure, or forming a group of groups according to a pre-specified number. This process includes a series of steps, ranging from Pre-processing and development of algorithms, up to the validity of the solution and evaluation. Each of them is closely related to each other and it poses great challenges to scientific disciplines. Here we focus on grouping algorithms and revise a broader one Variety of approaches that appear in the literature. These algorithms evolving from different research communities, we intend to solve different problems, and have their advantages and disadvantages. Though we have already seen many examples of successful applications of conglomerate analysis, there are still many outstanding problems due to the existence of many intrinsic insecure factors. These problems they have already attracted and will continue to attract intense efforts of wide disciplines.

REFERENCES

[1] J.-T. Chien, "Hierarchical theme and topic modeling," IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 3, pp. 565–578, 2016.

[2] Bernardini, C. Carpineto, and M. D'Amico, "Full-subtopic retrieval with keyphrase-based search results clustering," in IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intelligent Agent Technol., 2009, pp. 206–213.

[3] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, "Self-organization of a massive document collection,"IEEE Trans. Neural Netw., vol. 11, no. 3, pp. 574–585, 2000.

[4] K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram, "A hierarchical monothetic document clustering algorithm for summarization and browsing search results," in Proc. Int. Conf. World Wide Web, 2004, pp. 658–665.

[5] R. Xu and D. Wunsch, "Survey of clustering algorithms," IEEE Trans. Neural Netw., vol. 16, no. 3, pp. 645–678, 2005.