Academic Journal of Computing & Information Science
ISSN 2616-5775 Vol. 1, Issue 1: 28-35, DOI: 10.25236/AJCIS.010004

Analysis of Massive Unstructured Data Model Based on Clustering Algorithm

Yuxiang Cai, Lijun Cai, Ting Fu,Yong Ye, Sheng Zhou

State Grid Fujian Electric Power Co., Ltd. information communication branch,Fuzhou,Fujian 350003,China

ABSTRACT. *With the rapid development of the Internet and the Internet of Things, the degree of informatization of various industries has rapidly increased, and modern computers can easily collect large amounts of data. Therefore, various industries have begun to adopt large-scale databases to collect more. The information, and through this information to get more knowledge, resulting in a huge amount of data. In this paper, the clustering algorithm is used to analyze massive unstructured data, and the parameters such as the number of accesses and the duration of use are considered, so that the unstructured data can improve the accuracy when searching, so as to more effectively meet the needs of users.*

KEYWORDS: *Clustering algorithm; Unstructured data; Data clustering; Clustering evaluation; Data mining*

1. Introduction

The size of the data on the Internet has reached the level of TB. On the one hand, people feel that the scale of this data is too large, making people feel unpredictable and unable to start. On the other hand, people also realize that there are many very meaningful and valuable information hidden behind such large-scale data [1]. For example, the statistics of weather information generated by satellite remote sensors, the prediction of market information in the financial and stock market, the monitoring of network text or media information, and the recording of commodity information. These data have common characteristics, not only massive, but also high-speed, real-time arrival. This uninterrupted data is called stream data and can also be called data stream. The data stream has now become the main way of data existence, and it has gradually been valued by people. Data streams are very different from traditional data. Unlike traditional data sets, data streams continuously arrive and leave the computer system at ever-changing rates. These data are real-time, fast, massive, and potentially unlimited.

With the increasing emphasis on big data applications, many data mining clustering algorithms have emerged in the data mining field of big data, including clustering algorithms based on data density, clustering algorithms based on hierarchical analysis, and

data type partitioning. Clustering algorithms, etc., all kinds of clustering algorithms have their distinctive features, and have a good application effect in a certain field of data mining. For example, the BRICH algorithm is highly efficient and suitable for convex or spherical clustering types. Insensitive to noise and input data; DBSCAN algorithm is generally efficient, suitable for clustering of arbitrary shapes, sensitive to noise and data input; CURE algorithm is more efficient, suitable for any shape clustering type, for noise And the input data is not sensitive; CLARANS algorithm is less efficient, suitable for convex or spherical clustering type, sensitive to noise and sensitive to input data; CLIQUE algorithm is inefficient, suitable for convex or spherical clustering type, Noise and input data are less sensitive; K-Means algorithm is less efficient, suitable for convex or spherical clustering types, less sensitive to noise and input data Wait.

2. Unstructured Data

2.1 Unstructured data introduction

In recent years, with the rise of technologies and applications such as the mobile Internet, the Internet of Things, and social networks, the amount of data worldwide has grown rapidly. According to IDC's research report, by 2020, global data usage is expected to surge 44 times, reaching 35.2ZB, which means that approximately 37.6 billion 1TB hard drives are needed to store data [2].

Data is structured, semi-structured, and unstructured. Structured data refers to data that can be expressed in two-dimensional relationships; semi-structured data refers to data with certain structure such as XML and web pages; non-structured data has a fixed data structure and cannot be stored in relational databases. Stored in various types of files, it covers a variety of data types, including office documents, reports, corporate logs, customer service/chat logs, emails, doctor's diagnostics, images, and audio/video. According to statistics, 20% of the data in the enterprise is structured, and 80% is unstructured or semi-structured. The technology of structured data retrieval has matured, and enterprises can carry out in-depth mining and then react decision-making information to enterprises, and enterprises benefit from it. On the other hand, the unstructured data, which accounts for a considerable proportion, can only be shelved, and few people care about it. Even a diamond-like value is hard to show its light. In order to make decisions, in addition to analyzing internal information, external data is more essential, and these external data are unstructured data for enterprises. Data shows that 58% of corporate executives rely on unstructured data analysis when making business decisions, and this number will increase as the level of informatization increases [3]. It can be seen how important unstructured data is to the enterprise.

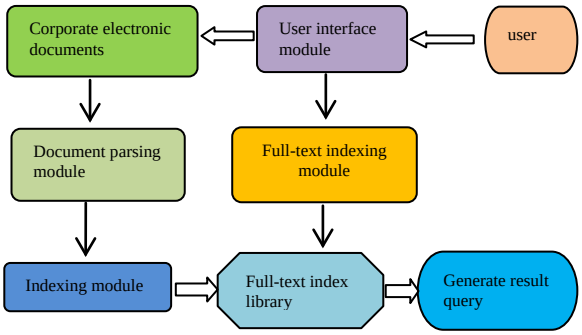

Figure 1 Unstructured data full-text analysis process

2.2 Unstructured data analysis

Unstructured data management usually includes modeling, storage, retrieval, analysis, application and other aspects. The full-text search flow chart of the enterprise is shown in Figure 1. These unstructured data, such as PPT, flash files, audio and video, etc., are created by employees themselves, as well as emails from partners, as well as downloaded from the Internet. Unstructured data is accumulating more and more, and you want to search for the files you need, sometimes you need to go through multiple searches to find them, and sometimes you need to manually find them one by one, which is very time consuming. The most common method we use in the search process is to use keyword search. Its advantage is that the keywords are set by themselves, which is in line with the habits of ordinary individuals, but it also has certain drawbacks. The following is an example of a pregnancy check-up item in a hospital system. If you want to have all the documents related to the "pre-pregnancy check", you need to enter the keyword "pre-pregnancy check document" to search, as long as the file name contains "pre-pregnancy check documents". Similar text, then it will appear in the search results. However, it is also possible that the keywords are not completely matched, so that the more important data and the data that the user needs are not searched. Moreover, the user's thoughts cannot be understood only by keyword search, so the search results provided are highly likely to have low accuracy, which makes the user unsatisfied. In order to improve the efficiency and accuracy of unstructured data, and to meet the needs of users, we propose to use clustering algorithm to analyze and solve this problem.

3. Clustering Algorithm Applied To Unstructured Data Analysis

Data stream clustering is a new technology that has gradually emerged in recent years. As mentioned earlier, data streams have many features that are not available in static data sets, and these features increase the difficulty of clustering data. Therefore, in recent years, many domestic and foreign scholars have studied this. Based on the traditional classical algorithms, many improvements and extensions have been proposed to adapt to this

dynamic data flow. Traditional clustering algorithms can be roughly divided into five types: partition-based methods, hierarchical-based methods, density-based methods, grid-based methods, and model-based methods. The data stream clustering method can also be roughly distinguished in this way.

Based on the partitioning method, the data is built into k basic clusters, and then the data is divided into the nearest cluster according to the distance between the input data and the k clusters, and the cluster is adjusted. In the traditional clustering algorithm, the typical representative based on the partitioning method is the K-Means algorithm and the K-Medians algorithm [4].

Guha et al. proposed a new algorithm for single pass scanning to process data streams. The algorithm is based on the K-Means algorithm and has a low demand for storage space. Suppose the amount of data in the data stream is N, and the number of center points to be clustered is K. In addition, a constant ε less than 1 is needed in the algorithm. The time complexity of this algorithm is O(nk) and the space complexity is O(n ε). However, there is an approximate intermediate result in this algorithm. This intermediate variable will become larger as the number of iterations increases, which will eventually lead to inaccurate clustering results. Charikar et al. improved the algorithm, and the improved algorithm used a divide-and-conquer strategy to further optimize the algorithm.

O'Callaghan et al. proposed the STREAM algorithm as a representative algorithm in the extended partitioning method. STREAM is derived by extending the K-Medians algorithm. The algorithm also uses a single pass scan and uses a constant factor with a time complexity of O(kn) and a spatial complexity of O(N ε). In this algorithm, m buckets are set, then the points in each bucket are divided into k clusters, and the information of the bucket is represented by the information of the center points of the k clusters. After that, the data points used for statistics are discarded, and only the information of the center point is retained. When a certain number of center points are collected, these points are clustered again, and the clustering process is repeated. In the STREAM algorithm, k is not a global static constant, but is dynamically changed, and the size tends to k only at the end of the algorithm.

3.1 Definition of clustering algorithm

Clustering is the process by which objects are aggregated and form groups of objects (classes) or clusters with similarities. That is, the generated class is a set of data objects, the objects in the same class are similar, the objects in different classes are different, and the data distribution characteristics in the data set can be found through clustering. How to obtain useful knowledge from people in the data stream is a very difficult challenge. The traditional data processing method cannot adapt to this new type of data, and new technologies are urgently needed to realize it. Clustering algorithms are used to study unstructured data. Data stream mining is a new data mining technology that has emerged in recent years. Different from traditional data mining, data stream mining is based on the characteristics of data streams [5].

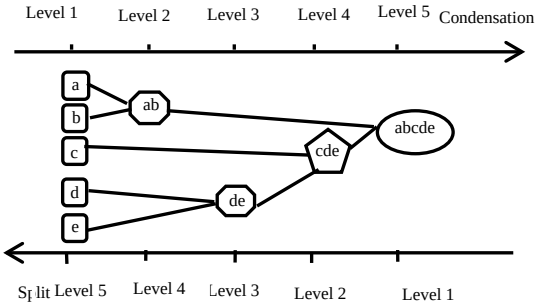

Figure 2 Schematic diagram of clustering algorithm based on hierarchy

Clustering is a very important direction in traditional data mining technology, and it also plays a pivotal role in data stream mining. For the clustering process, Han and Kamber have such a definition in the literature: "The clustering process divides data objects into regions of different sizes. The data in the same region is highly similar, and the data in different regions are highly different. In this way, data objects are divided into classes or clusters." Data stream clustering can be applied to network intrusion detection, weather monitoring, emergency response systems, stock trading, e-commerce, telecommunications, planetary remote sensing, website analysis, etc. It is an important subject, and it is a very important new field and new technology.

3.2 Description of the clustering algorithm

The hierarchy-based approach is to decompose or merge data objects according to their distances, usually in two ways: splitting and condensing. The way of splitting is top-down, starting with all the data in a cluster, and then gradually breaking down the data into multiple clusters through loop iteration; the way of condensing is just the opposite, starting from the bottom up, starting with each data divided into one Clusters, then clusters of similar distance will be merged into one. In the traditional clustering method,

The typical representative of the hierarchical method is the BIRCH algorithm. A schematic based on the hierarchical approach is shown in Figure 2.

Aggarwal et al. proposed the CluStream algorithm, which is a typical algorithm for extending the hierarchy. In fact, the CluStream algorithm is not just an algorithm, but an algorithm framework. It breaks down the entire process of data stream clustering into two steps: first, online updates, then offline clustering. The main work of the online update phase is to collect data in the data stream and update the micro-cluster. The algorithm uses the pyramid time frame structure to store the data, and then updates the micro-cluster. The main task of the offline clustering phase is to macro cluster the micro-cluster. Class, the clustering method here extends the traditional BIRCH algorithm. The CluStream algorithm is one of the more popular data stream clustering algorithms, and the two-stage process

proposed by the CluStream algorithm is now widely used in various data stream mining algorithms.

The HPStream algorithm was proposed in Aggarwal et al. This algorithm is aimed at clustering high-dimensional data streams and improves the CluStream algorithm. By using the structure of fading cluster and project clustering technology, better clustering of high-dimensional data streams is obtained. In the current popular data stream clustering algorithm, two algorithms, CluStream and HPStream, are recognized and used. Udommanetanakit et al. proposed the E-Stream algorithm, which was supplemented and improved based on the HPStream algorithm.

3.3 Application of clustering algorithm

After the data M is preprocessed, it retains d attributes, that is, the feature vector of the data is

$$\mathrm{D}\left(\mathrm{x}_1, x_2, ..., x_d\right)$$

Then map it to the d-dimensional grid space

$$\mathrm{g}\left(\mathrm{s}_1, s_2, ...s_d\right)$$

in. The mapping method is to set a mapping function for each dimension. Assume

$$\mathrm{func}_{\mathrm{map}}\left(\mathrm{s}_i\right) = x_i \bmod m_i \tag{1}$$

Where mi is the grid width of the i-th dimension.

Use this function to find the grid to which the data should be mapped, and record the eigenvectors of the grid. After vecpost, the grid update operation is performed.

Defining the mesh density Den can be expressed as Den(t), where t is the most recent data arrival time. It is specified that when t = 0, Den(t) = 0. This rule guarantees that the first mapped data will not be affected by the attenuation in each grid. Also, assume that the data in the grid arrives at time t1, t2, ..., tk at time t. Then there is

$$Den\left(t\right) = \lambda^{t^1} + \lambda^{t^2} + ... + \lambda^{t^k} \tag{2}$$

It can be proved that the grid updates the data once at time t, and if t + 1 is updated again, the mesh density at time t + 1 is

$$Den\left(t+1\right) = \lambda Den\left(t\right) + 1 \tag{3}$$

From the formula for calculating the mesh density, you can get:

$$\begin{aligned} Den\left(t+1\right) &= \lambda^{t_1+1} + \lambda^{t_2+1} + ... + \lambda^{t_k+1} + 2^0 \\ &= \lambda\left(\lambda^{t_1} + \lambda^{t_2} + ... + \lambda^{t_k}\right) + 1 \end{aligned}$$

$$= \lambda Den(t) + 1 \quad (4)$$

This property only applies to data that arrives in the same grid without interruption for continuous time. Assume that the data arrives at the grid for the last time at time ta, after which the grid remains unchanged, no new data arrives, and new data is mapped to this grid until tb.

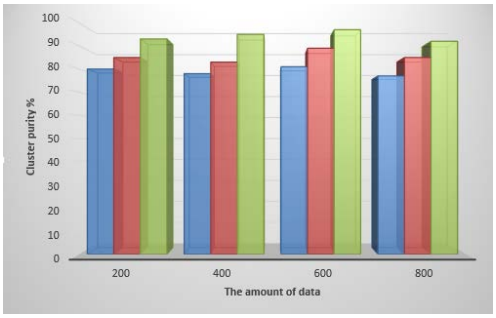

Figure 3 The effect of window size on clustering accuracy

The following experiment is to use clustering algorithm to cluster some data sets and compare the cluster purity with different data volume and grid size. Here, the attenuation factor λ = 0.98 is set. For the four cases where the data volume is 1 * 105, 2 * 105, 3 * 105, 4 * 105, a total of 4 experiments were performed, setting the number of meshes per dimension to 20, 40, and 60 respectively. The cluster purity results are shown in Figure 3. It can be clearly seen from this experiment that the more the number of window divisions, the higher the purity of the cluster. When each dimension contains only 20 grids, the average cluster purity is about 80%; when each dimension contains 40 grids, the average cluster purity is increased to more than 85%, and the purity is improved. 5%; when each dimension contains 60 grids, the average cluster purity is about 95%, which is nearly 10% higher, and contains very few non-similar data. After comparing the labels of the datasets, most of the non-similar data are at the edge of the grid group, or the data with higher attenuation and longer history.

However, it should be noted that in the clustering algorithm, the speed of the offline clustering part is related to the number of non-empty grids. In the operations of grid updating and merging grid groups, the calculation time is both right and wrong. The number of empty grids is directly related. When the number of meshes increases, the meshing granularity is finer, and the data is also scattered and mapped to many different meshes, which results in a slower statistical non-empty mesh, and the performance of the clustering algorithm is obvious. Considering the purity of the cluster, the problem of clustering speed should also be considered. In the actual application scenario, the data stream usually changes at a high speed. For the data stream clustering algorithm, a faster response speed is required. For the clustering algorithm, the number of grids is required to be limited. Therefore, the setting of the grid size also needs to be compromised.

4. Conclusion

Clustering algorithm is a very important research direction in data mining. The basic idea of the clustering algorithm classifies data similar to each other into the same class according to similarity, and data different from each other does not belong to the same class. Cluster analysis can be used to discover the distribution of data, the relationship between data attributes, etc., and even clustering as a data preprocessing link in data mining. Clustering algorithms are used to classify customers, find marine areas that have an impact on the Earth's climate, and compress data. In this paper, the clustering algorithm is used as the research method to analyze the effective search of massive unstructured data, and consider the parameters such as the number of visits and the duration of use, so that the unstructured data can improve the accuracy when searching for applications, thus more effectively satisfying users.

References

[1] Yin, C., Zhang, S., Xi, J., & Wang, J. (2016). An improved anonymity model for big data security based on clustering algorithm. Concurrency & Computation Practice & Experience, Vol. 29, No.7, pp.11-19.

[2] Liu, Sanya Ni, Cheng Liu, Zhi Peng, Xian Cheng, Hercy N. H. (2017). Mining individual learning topics in course reviews based on author topic model. International Journal of Distance Education Technologies, Vol. 15, No.3, pp.118-126.

[3] GJ Edgar, AE Bates, TJ Bird, AH Jones, S Kininmonth, & RD StuartSmith, et al. (2016). New approaches to marine conservation through the scaling up of ecological data. Ann Rev Mar Sci, Vol. 8, No.1, pp.435-461.

[4] Quantin, M., Hervy, B., Laroche, F., & Bernard, A. (2016). Supervised process of un-structured data analysis for knowledge chaining . Procedia Cirp, No. 50, 436-441.

[5] Leng, J., & Jiang, P. (2017). Mining and matching relationships from interaction contexts in a social manufacturing paradigm. IEEE Transactions on Systems Man & Cybernetics Systems, No. 99, pp.11-13.