INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 3)
Available online at: www.ijariit.com

uDCLUST: A novel algorithm for clustering unstructured data

Aamir Ahmad Khandy
khanday.aamir@gmail.com
Dr. C.V. Raman University, Kargi Road Kota, Bilaspur Chittisgrah

Dr. Rohit Miri
rohitmiri@gmail.com
Dr. C.V. Raman University, Kargi Road Kota, Bilaspur Chittisgrah

ABSTRACT

Data that has been arranged and systematized into an organized and formatted repository, usually a database, so that its elements and essential features and can be made directly accessible for more powerful and adequate processing and analysis is known as Structured Data. Un-structured data is data that doesn't fit accurately in a traditional database and has no identifiable internal structure and a predefined data model. We cannot perform different operations like update, insert and delete on unstructured data. Clustering is a process of unsupervised learning and is the most common method for mathematical and demographic data analysis. It is the main task of preliminary data mining, and an ordinary technique for statistical data analysis, mathematical data analysis, demographic data analysis, used in many fields, including ML (Machine Learning), recognition of patterns, analysis of images, retrieval of information, bioinformatics, compression of data and computer graphics. Available clustering algorithms have the difficulty to determine the number of clusters in a dataset and also are difficult to cluster outliers even that have common groups. A final related drawback arises from the shape of the data cluster where it is difficult and complex to cluster non-spherical and overlapping datasets. In this framework, we intended and designed an algorithm called uDCLUST (Un-structured Data Clustering), which identifies an appropriate number of clusters in unstructured data as well as cluster outliers easily with non-spherical and overlapping datasets.

Keywords— *Big data, Unstructured data, Clustering algorithms, MongoDB*

1. INTRODUCTION

Big Data has become one of the buzzwords in Information Technology and computer science during the last couple of years. In the current digital era, according to massive progress and development of the internet and online world technologies such as big and powerful data servers, we face a huge volume of information. Data size has increased dramatically with the advent of today's technology in many sectors such as manufacturing, business, science and web application. With the rising of data sharing websites, such as Facebook, Flicker, YouTube, Twitter, Google, Google news, there is a dramatic growth in the number of data.

Unstructured data (or unstructured information) is information that either does not have a pre-defined data model or is not organized in a pre-defined manner. It may be typically human generated such as Text files, Email, Social Media, Website, Mobile data, Communication, Media; and machine-generated including satellite imagery, scientific data, digital surveillance, sensor data etc. Unstructured data is not normalized and structured but is stored in easily accessible and joined formats. These formats make it not difficult to give or exchange information. Unfortunately, that lack of difficulty also makes unstructured data open to attack to unauthorized access.

Structured Data that is arranged and systematized into an organized and formatted repository, usually a database, so that its elements and essential features and can be made directly accessible for more powerful and adequate processing and analysis.

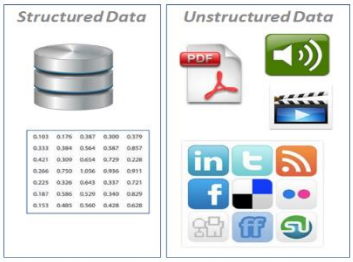

Fig. 1: Structured and Unstructured data [7]

Clustering is a process of grouping data objects into disjointed clusters so that the data in the same cluster are similar, but data belonging to different cluster differ. A cluster is collections of a data object that are similar to one another are in the same cluster and dissimilar to the objects are in other clusters. The appeal for organizing the sudden and rapidly increasing data and learning relevant and profitable information from data, which makes clustering techniques that are widely correlated in many application areas such as AI (Artificial Intelligence), biology, CRM (Customer Relationship Management), DS (Data Compression), DM (Data Mining), IR (Information Retrieval),

IP (Image Processing), ML (Machine Learning), marketing, medicine, pattern recognition, psychology, statistics and so on. Cluster analysis is a tool that is used to observe the characteristics of the cluster and to focus on a particular cluster for further analysis. Clustering is unsupervised learning and does not rely on predefined classes.

Clustering is the most valuable unsupervised learning problem that is to be considered. Like every other difficulty of this kind, it deals with finding a relationship or structure in the collection of unlabelled data. Thus a cluster is, therefore, a group of objects which are very much alike between them and are dissimilar to the association of the object to other clusters. We can show this with a simple graphical example (refer to figure-2).

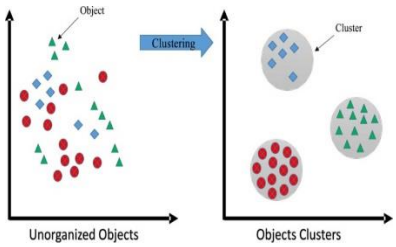

Fig. 2: Example of clustering

1.1 Challenges of clustering

A cluster is a collection of similar documents, and clustering is also called unsupervised learning conducted for the association of data on the basis of some similarity measure, accordingly without having to pre-specify classification. We don't have any instruction data to develop a classifier that has learned to categorize data. Without any earlier knowledge of the number of categories, group size, and the type of information i.e. unstructured information, the problem of clustering come into sight challenges. In-use algorithms suffer from these problems,

- Supports only a small size of datasets.
- It is very difficult to organize the outliers even if that data contains a common group.
- It is also complex and hard to determine in overall the total number of clusters in a dataset.
- It is difficult and very complex to organize and cluster non-spherical data as well as the data that is overlapping.
- Current clustering techniques do not address all the requirements adequately and concurrently.
- Dealing with a large number of dimensions and a large number of data items can be problematic because of time complexity.
- A new similarity model is needed for unstructured data as traditional distance functions cannot capture the pattern similarity among the objects.

1.2 Our contribution

In day-to-day life, most of the real-time data are unstructured in nature. There are no direct algorithms available to cluster these unstructured data. Unstructured data can be transformed into a well-defined schema and later the available clustering algorithms can be applied on these structured data. uDCLUST aids in clustering the unstructured data directly. The following advantages are achieved by using uDCLUST:

- Preprocessing of unstructured data is avoided and
- Time complexity is somehow a lit bit reduced.
- Direct approach in organizing un-structured data.
- Overhead of distance functions is reduced.
- This is the simplest way to put data in a Relation Database Management (RDBMS) where we can Query (to perform different operations) data.
- We can cluster the outliers contained in dataset even if that data contains a common group.
- This method can determine in overall the total number of clusters in a dataset easily.
- This method can organize and cluster non-spherical data as well as the data that is overlapping easily.

These Algorithm clusters un-structured data with only a single attribute very fast and effectively. But, when organizing the unstructured data with more than one attributes, the time complexity is increased while processing data that has a large size.

2. RELATED WORK

Work performed in 2018 [28] presented about role of Machine Learning and Big Data Processing and Analytics (BDA). The development of Machine Learning and Big Data Analytics is complementary to each other. He discussed various future tends of Machine learning for Big data. Data Mining implies how Machine Learning can be made more intelligent to acquire text or data awareness.

(A work performed in 2017[21] presented new ways of processing Big Data through Machine Learning Algorithms. Due to Big Data characteristics, traditional tools are now not capable of handling its storage, transport or its efficiency.

Work performed in 1996[32] stated in "An Efficient Data Clustering Method for Very Large Databases ", Birch (Balanced Iterative Reducing and Clustering using Hierarchies), and demonstrates that it is especially suitable for very large databases. BIRCH incrementally and dynamically clusters incoming multi-dimensional metric data points to try to produce the best quality clustering with the available resources (i. e., available memory and time constraints).

Work performed in 2014[4] stated in "DMM-Stream: A Density Mini-Micro Clustering Algorithm for Evolving Data Streams "Clustering real-time stream data is an important and challenging problem. The existing algorithms have not considered the distribution of data inside micro cluster, specifically when data points are non-uniformly distributed inside the micro cluster. In this situation, a large radius of the micro cluster has to be considered which leads to lower quality. In this paper, they have proposed DenStream, an effective and efficient method for clustering an evolving data stream.

Work performed in 2014[9] proposes a classification algorithm for Big Data based on feature selection. Firstly, the feature selection algorithm is designed to reduce the size of the dataset. Then, a parallel k-means algorithm is applied to the data subsets selected in the first step. Experimental results show that the proposed algorithm provides better classification accuracy than existing algorithms and takes much less time than other classification algorithms for Big Data.

Work performed in 2012[19] proposes a classification algorithm and feature extraction for Big Data that is based on PCA and LS-SVM, the experimental results on the Big Data shows that the proposed classification algorithm based on feature extraction allows solving large classification problems.

All these problems made us develop a tool comprising of an algorithm to cluster the unstructured data. Most of the big data are unstructured data; clustering these unstructured data becomes a challenging task. So, we designed and developed a direct method for normalizing the unstructured data. We tested our algorithm with 5 different datasets.

3. TOOLS USED IN THIS ALGORITHM

Figure-3 shows the architecture design of this Algorithm. We are going to design two Web Portals using PHP (Hypertext Preprocessor) to collect unstructured information (Aadhaar, Election Commission) from the user and stored in MongoDB database that is efficient in storing unstructured data). The unstructured data that is stored in the MongoDB database are processed and converted to JSON (JavaScript Object Notation) documents. This algorithm collects this JSON (JavaScript Object Notation) documents and does the clustering process by using a singly linked list that is linear data structure and dynamic data structure. This algorithm then generates the output of clustered information and results for the unstructured data that is based on the input key. The design of the proposed algorithm is the organization of web portal, MongoDB, JSON and Linked List. First, we have to create two web portals through PHP to collect big data that is in unstructured format i.e. Aadhaar and Election data. The interface of the web portal is designed using PHP programming language. This data is gathered and stored in MongoDB efficiently. After the data is stored in MongoDB, it is then processed using a JSON object. This JSON object is then processed to generate clustered output.

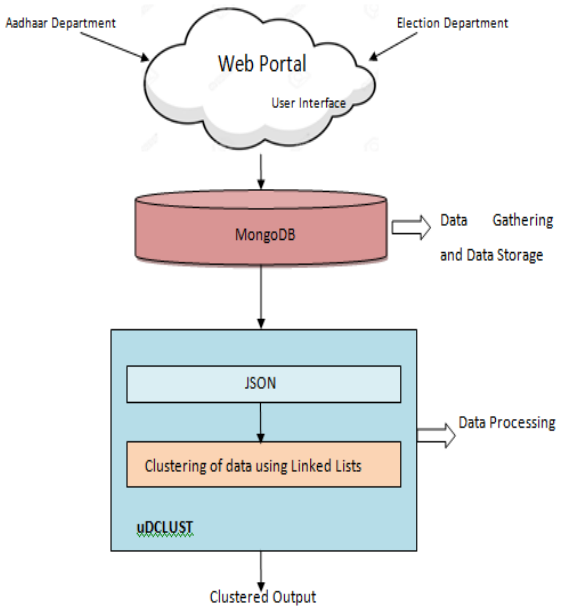

Fig. 3: Design of algorithm

3.1 Algorithm

Input:

U= {n1, n2, n3, ……n} // Set of n number of data points

Output:

A set of R Clusters. // Number of the desired Cluster.

Steps:

Step 1: Read JSON document for clustering attribute;
Step 2: Read each record for key availability until the End of File;
Step 3: If the key is available then extract a unique identification key for each record present in a particular document.
Step 4: Extract all the attributes and their values specified with them for each particular specified key.
Step 5: Attach all the keys/values to a dynamic linked list.
Step 6: Traverse each record/node that we have stored in the linked list.
Step 7: If the new distinct node is found then form a new list and delete the node from the existing list.
Step 8: Else attach this particular node to the existing list already stored and delete the node from the existing list.
Step 9: Else if the key is not available then go to step 2;
Step 10: Output all the linked lists created;

3.2 Working of the proposed algorithm

Figure-4 shows the working of this algorithm. The Aadhaar and Election Commission datasets are stored in MongoDB. Each user is identified and detected using a key. All the attributes of a user are mapped to the key using a hash method that is efficient. The un-structured datasets that are stored in the MongoDB database are processed and converted to JSON (JavaScript Object Notation) documents. This conversion is done using different tools which is available in MongoDB. The JSON documents are then forwarded to the desired algorithm for clustering. The clustering attribute is got as input from the user. The clustering process is done using the dynamically linked lists. For each new different value found in the JSON (JavaScript Object Notation) document, a new linked list is created. These linked lists created to represent the clusters or groups.

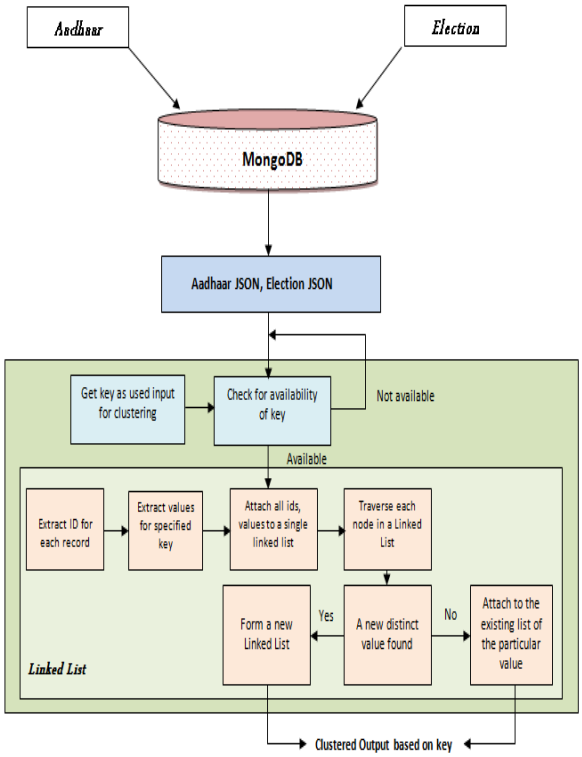

Fig. 4: Working of the proposed algorithm

3.3 Example

A model of JSON (JavaScript Object Notation) document that is based on the detail of a company with employees working on different projects along with their gender is assembled based on the attributes Project name and country. Figure-4 shows the assembled output based on only a single attribute and also the Figure-5 depicts the organized output based on two attributes i.e. project name and country.

3.3.1 Model of JSON (JavaScript Object Notation) file

```
[{
"Employee1":{"firstname":"sekar","country":"India","projects"
:"[\"Project-P3\",\"Project-P6\",\"Project-P8\"]"},
"Employee3":{"firstname":"Rani","lastname":"shankar",
"country":"China","projects":"[\"Project-P6\",\"Project-P3\"]"},
"Employee2":{"projects":"[\"Project-P3\",\"Project-P6\"]",
"country":"China", "firstname":"vijay","lastname":"sampath"}
}]
```

Clustering key - projects, country

3.3.2 Clustered output - single attribute (Projects)

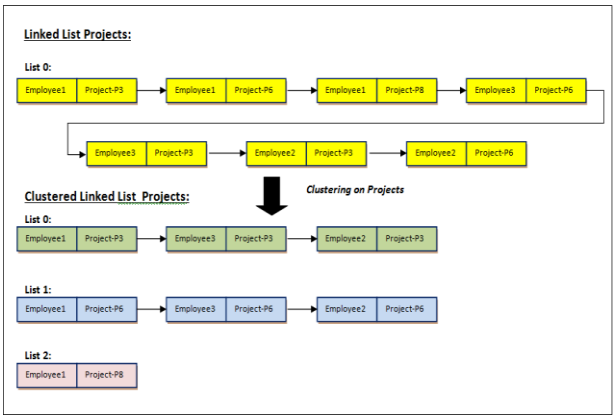

Fig. 5: Clustering linked list based on projects

3.3.3 Clustered output - two attribute (Projects and Country)

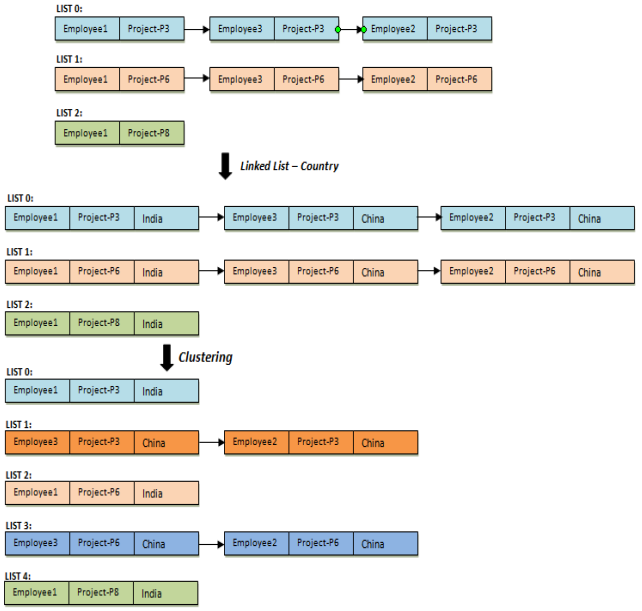

Fig. 6 Clustering linked list based on project and country

4. ENVIRONMENT SETUP AND DATA INPUT

We use Java language to implement our algorithm. Experiments were conducted on a machine consisting of operating system Windows 8.1 on Intel i7-4700MQ CPU, 8 GB RAM memory.

We used five datasets, out of five, three datasets are already available and they are classified in Table-1. The remaining two datasets, Aadhaar and Election commission are generated by using Web Portal which is designed using PHP.

Table 1: Details of the datasets.

Data Set	No of records	Format	Link
Facebook Social Media	3,00,000	JSON	https://data.facebook.us/api/views/5b3ars48/rows.json
Instagram Social Media	3,00,000	JSON	https://data.instagram.gov/api/views/rfcd-nnxe/rows.json
Twitter Social Media	3,00,000	JSON	https://data.twitter.gov/api/views/5ytf-wban/rows.json
Aadhaar	3,00,000	JSON	Manually created and Collected using Web Portal and stored in Redis
Election Commission	3,00,000	JSON	Manually created and Collected using Web Portal and stored in Redis

5. RESULTS

uDCLUST algorithm for clustering single attribute was implemented and tested for the datasets specified above. Table-2 shows the running time for clustering a single attribute of the datasets.

Table 2: Running time for single-attribute clustering.

Data Set	Running time for one attribute					
	1000 Records	5000 records	10000 records	50,000 records	1,50,000 records	3,00,000 records
	Seconds					
Aadhaar	3	5	8	12	16	24
Election Commission	4	7	11	14	21	29
Facebook Social Media	5	8	12	18	23	28
Instagram Social Media	3	5	8	12	16	24
Twitter Social Media	7	11	15	19	23	27

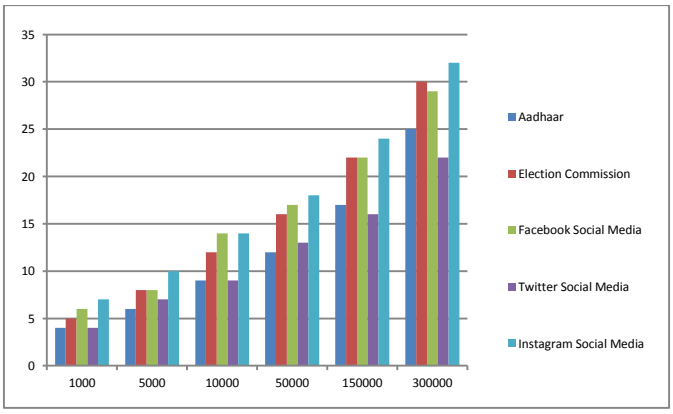

Fig. 7: Running time for one attribute clustering

Table 3: uDCLUST running time for clustering of two attributes

Data Set	Running Time for One Attribute					
	1000 records	5000 records	10000 Records	50,000 records	1,50,000 records	3,00,000 records
	Seconds					
Aadhaar	71	163	299	443	801	940
Election Commission	82	172	322	474	861	980
Facebook Social Media	87	176	334	480	871	988
Instagram Social Media	70	151	287	441	816	971
Twitter Social Media	92	214	344	483	877	1043

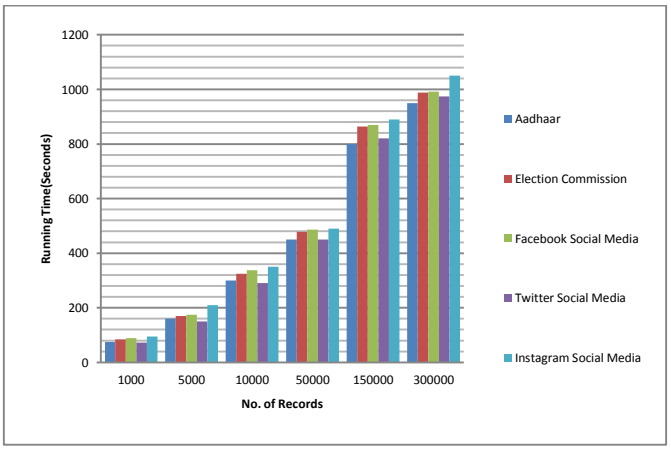

Fig. 8: Running time for the two attribute clustering

Table 4: Comparison of results for clustering

Data Set	3,00,000 records	
	Single Attribute	Two Attributes
Aadhaar	68	2717
Election Commission	86	2891
Facebook Social Media	94	2936
Instagram Social Media	68	2736
Twitter Social Media	102	3053

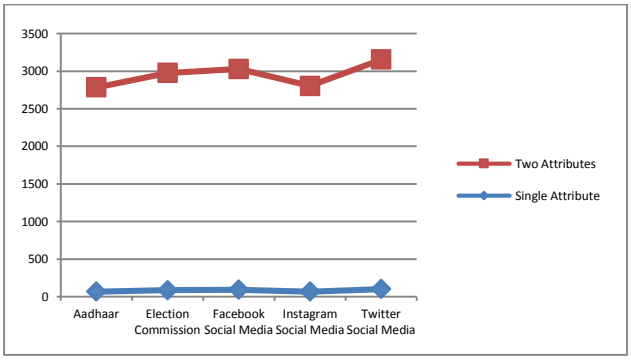

Fig. 9: Comparison of results for clustering 3, 00, 000 records

From table 4 and figure 9, it can be realized that when clustering a large data the processing time increases. To reduce this, the clustering process can be parallelized.

6. CONCLUSIONS AND FUTURE RESEARCH

Data that has been arranged and systematized into an organized and formatted repository, usually a database, so that its elements and essential features and can be made directly accessible for more powerful and adequate processing and analysis is known as Structured Data. Un-structured data is data that doesn't fit accurately in a traditional database and has no identifiable internal structure and a predefined data model. We cannot perform different operations like update, insert and delete on un-structured data. Clustering is a process of unsupervised learning and is the most common method for mathematical and demographic data analysis. It is the main task of preliminary data mining, and an ordinary technique for statistical data analysis, mathematical data analysis, demographic data analysis, used in many fields, including ML (Machine Learning), recognition of patterns, analysis of images, retrieval of information, bioinformatics, compression of data and computer graphics. Available clustering algorithms have the difficulty to determine the number of clusters in a dataset and also are difficult to cluster outliers even that have common groups. A final related drawback arises from the shape of the data cluster where it is difficult and complex to cluster non-spherical and overlapping datasets. In this framework, we intended and designed an algorithm called uDCLUST (Un-structured Data Clustering) algorithm, which identifies an appropriate number of clusters in unstructured data as well as cluster outliers easily with non-spherical and overlapping datasets. The projected algorithm is practised in 5 different datasets and results are classified and figured. We manually created two new dataset called Aadhaar and Election Commission data and stored it in a Redis database, which is a key-value store, stored in an unstructured format. The results obtained are tabulated and shows that the designed clustering algorithm works well, as every work may lag at some point, our uDCLUST algorithm lags while clustering more than 5 fields. The performance of the algorithm degrades in terms of compilation time; this problem can be solved by using the Parallel Computing and also the algorithm should be defined using Python programming language, which is our future research.

7. REFERENCES

[1] A. L. Mohammad, K. A. (2018). A hybrid strategy for krill herd algorithm with the harmony search algorithm to improve the data clustering. Intelligent Decision Technologies.

[2] Ahamed Shafeeq B M, H. K. (2012). Dynamic Clustering of Data with Modified K-Means Algorithm. International Conference on Information and Computer Networks vol 2, (pp. 221-225).

[3] Ahmad Amir, a. L. (2007). "A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering, 503-527.

[4] Amini, A. T. (2014). On density-based data streams clustering algorithms: A survey. Journal of Computer Science and Technology, 116-141.

[5] Banker. (2012). MongoDB in Action. Shelter Island: Manning Publications Co.
Barlow, H. (1989). "Unsupervised learning". Neural Computing, 295-311.

[6] Bazes Richardo, R. Y. (1999). Modern Information Retrieval. ACM Press.

[7] Brook. (2018, Dec 5). Structured vs Unstructured Data: How to Protect Your Organization's Data. Retrieved May 5, 2019, from DigitalGuardian: https://www.laserfiche.com/es/ecmblog/4-ways-to-manage-unstructured-data-with-ecm/

[8] Chen, J. (2017). Structured and Unstructured Data. Retrieved from Smart Data Collective: http://smartdatacollective.com/michelenemschoff/206391/quick-guide-structured-and-unstructured-data

[9] Chen, Q. Z., & Wang. (2014). Incomplete Big Data. In Digital Home (ICDH) (pp. 263-266). IEEE.

[10] Chim Hung, D. X. (2008). Efficient Phrase-Based Document Similarity for Clustering. Knowledge and Data Engineering, IEEE Transactions, 1217-1229.

[11] Dean, G. S. (2010). MapReduce: a flexible data processing tool. Communications of the ACM, 53, 72-77.

[12] Dhillon, I. S. (2003). A divisive information-theoretic feature clustering algorithm for text classification. Journal of machine learning research, 1265-1287.

[13] Everitt, B. (2011). Cluster analysis. Chichester, West Sussex, U.K: Wiley.

[14] Feldman, R., & Sanger, J. (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge Univ. Press.

[15] Ghemawat, G. H., & Leung. (2003). The Google File System. In Proceedings of Nineteenth ACM Symposium on Operating Systems Principles, SOSP (pp. 29-43). New York, USA: ACM.

[16] Google Search Statistics. (n.d.). Retrieved April 28, 2019, from Internet Live Stats: https://www.internetlivestats.com/google-search-statistics/

[17] Jain, C. R., & Rong. (2015, Feb 07). Clustering Big Data. Retrieved April 01, 2019, from Department of Computer Science and Engineering (University of Notre Dame): http://www.cse.nd.edu/Fu_Prize_Seminars/jain/slides.pdf

[18] K. A. Abdul Nazeer, M. P. (2009). Improving the Accuracy and Efficiency of the K-means Clustering Algorithm. Proceedings of the World Congress on Engineering.

[19] Kim. (2012). A Classifier for Big Data. In Convergence and Hybrid Information Technology, 505-512.

[20] Kramer, J. D. (2013). Planning Guide - Getting Started with Big Data. Intel IT Center.
[21] L'Heureux Alexandra, G. K. (2017). Machine Learning With Big Data: Challenges And Approaches. IEEE Access.
[22] Langseth Justin, V. N. (2007). Analysis and transformation tools for structured and unstructured data.
[23] Maier, M. (2013, Oct 13). Towards a Big Data Reference Architecture. Retrieved May 2, 2019, from Eindhoven University of Technology: http://www.win.tue.nl/~gfletche/Maier_MSc_thesis.pdf
[24] Malone, R. (2007). Structuring Unstructured Data. Forbes.
[25] Matti, P. (2010-2011). Introduction to store data in Redis, a persistent and fast key-value database. AMICT , 39-49.
[26] Maugis Cathy, C. G.-L. (2009). Variable Selection for Clustering with Gaussian Mixture Models. Biometrics, Vol: 65(3), 701-709.
[27] Ran Vijay Singh, M. B. (2011). Data Clustering with Modified K-means. International Conference on Recent Trends in Information (pp. 717-721). IEEE.
[28] Roheet, B. (2018). Machine Learning and Big Data Processing. A Technological Perspective and Review, 933-1034.
[29] Safeer Yasir, A. M. (2010). Clustering Unstructured Data (Flat Files), an Implementation in Text Mining Tool. International Journal of Computer Science and Information Security, 174-180.
[30] Shi Na, L. X. (2010). Research on k-means Clustering Algorithm. Third International Symposium on Intelligent Information (pp. 63-67). IEEE.
[31] Statistics. (n.d.). Retrieved from YouTube: https://www.youtube.com/yt/press/statistics.html
[32] Tian Zhang, R. R. (1996). BIRCH: an efficient data clustering method for very large databases. ACM Sigmod Record. Vol. 25.
[33] Twitter Statistics. (n.d.). Retrieved from Internet Live Stats: https://www.internetlivestats.com/twitter-statistics/
[34] Zhu Lin, C. F.-l. (2009). Generalized Fuzzy C-means Clustering Algorithm with Improved Fuzzy Partitions. IEEE Transactions, 578-591.

