Special Issue on Applied Economics and Finance

Clustering Assessment Tendency for Big Data Analytics: Extract Useful Knowledge[1]

Soraya Sedkaoui, University of Khemis Miliana, Algeria / SRY-Consulting, France, soraya.sedkaoui@gmail.com
Salim Moualdi, Dept of Economy, University of Khemis Miliana, Algeria, moualdis@yahoo.com

Abstract: The clustering method is one of the important methods that can be used to analyze the big volume of data that should be grouped accordingly as much as possible. Depending on the characteristics of the data available today and to deal with big data challenges, several clustering methods have been developed. But, in many situations, we cannot know a priori the number of clusters in the data set. This refers to an important problem in cluster analysis or determining the numbers of clusters. In this context, this paper describes some clustering methods, with special attention to the Visual Assessment Tendency (VAT) algorithm as one of the known methods. This algorithm is implemented in advanced technologies to analyze big data. More generally, VAT algorithms are important before applying clustering analysis to classify unlabelled data. Clustering analysis requires the use of visual presentations to help understand what distinguishes classes.

Keywords: Big Data, Clustering Tendency, K-means, Knowledge, Visual Sssessment Algorithm

1. Introduction

The analysis process of the quantities of data available today, or big data, is an important and complex process. This process focuses on the extraction of meaningful insights through the exploration of a large volume of data generated in real-time and in different formats (structured and unstructured data). This process introduces several steps, from data collection to deployment of the results. Each step contains a set of different techniques and algorithms that can be used to improve the results of the analysis process (Sedkaoui, 2018a).

But, a significant question to which the data analysis process must respond is to know how to organize the quantities of data, or the observation, into comprehensives forms and meaningful structures or taxonomies. To do so, some methods need to be developed and redesigned, in order to deal with big data challenges. In this context, Cluster analysis is one popular approach for responding such question. It presents a very important technique, especially in a big data context.

This unsupervised algorithm is used to discover the hidden structure of the data and separate the data into different homogeneous groups or clusters. This method is widely used in extracting knowledge from data. Depending on the criteria and the need for the analysis process, this method is basically applied to divide data into a number of groups. It can be used in data mining, text mining, pattern recognition, and image segmentation, etc.

Researchers and experts have developed and proposed many cluster algorithms (Hastie et al., 2013). These algorithms can integrate a priori domain knowledge with the cluster analysis process to better guide the method and improve the quality of the partitions. In big data age, we can use many clustering algorithms, such as K-means, Hierarchical model, Gaussian mixture methods, etc. The clustering method aims in one side to define the number of groups (or cluster) in the data set and determine, on the other side, the structures of the cluster.

It is evident that if we know *K* we can use one of the several clustering algorithms in order to find *K* groups. So, to qualify a k group, many indices are available. But in the case *K* is not known, we need to identify it, i.e. define firstly, the number of groups (or *K*) before thinking about the partition of the data set into *K* groups or the quality (validation) of the obtained partition.

Therefore, identifying the number of clusters, or what is called 'Assessment of clustering tendency', is an important thing to better achieve clustering analysis. Traditional clustering methods suppose a number of groups and thereafter attempt to determine properly the possible structure associated with these clusters. In the data revolution context, it is difficult to define this number, because the volume of the data is very large and its nature is complex.

For a large volume of data, characterized by its complexity and variety of formats, it is not easy to determine the K number of classes. To address this issue, researchers have developed some methods. These methods include the Visual Assessment of Tendency (VAT) algorithm, developed by Bezdek and Hathaway (2002).

The purpose of this paper is to highlight the importance of this method (clustering method) and show that, in the big data area, analytics methods must be adapted and operationalized to deal with the challenges imposed by this phenomenon in order to generate useful knowledge and make them operational. And one way to overcome these challenges is to have a big data cluster in a compact format that will still be an informative version of the entire data and capable to generate useful knowledge.

This is to say, that the aim of our study is to illustrate the importance of cluster analysis, especially VAT algorithm, in data exploration process to generate value, and the need to develop this kind of methods. The objective is to overview

[1] Bu çalışma Soraya Sedkaoui ve Salim Moualdi özet bildirisi olarak, "*ICOAEF'18, IV. International Conference on Applied Economics and Finance & EXTENDED WITH SOCIAL SCIENCES, November 28 – 29 – 30, 2018 / Kuşadası – Turkey*" Kongresinde sanal oturumda sunulmuştur ve kongre procedia özet kitapçığında basılmıştır.

the techniques adopted to cluster the large amounts of data and to highlight strategies for data analysis process in order to extract value from it.

2. Challenges imposed by big data

A large volume of data is produced and collected every day. This volume, which is now measured in petabytes and zettabytes, is generated thanks to the advent of digital technology and smart connected devices. People and machine are now producing more and more data. Many enterprises collect and store data on their client's behavior. As such, many processes are being monitored by the machine (Sedkaoui and Gottinger, 2017). Experts, over the world, have defined many big data challenges related to its three characteristics usually used to define the big data phenomenon, i.e. the volume, variety, and velocity of data.

The volume, or the quantity of data, is growing exponentially. Facing this volume generated in real-time and coming from several sources and in many formats, explore the data to generate patterns and value become so complex. The process of data analysis depends now on the nature of data, their characteristics, the capacity of storage and treatment, technology infrastructure, the quality of data, its complexity, heterogeneity, scalability, and security … in addition to the need for advanced analytics techniques to improve results.

So, the various challenges of big data and its applications relating within society-at-large have been widely discussed in the literature (Boyd and Crawford, 2012; Ekbia et al., 2015; Sedkaoui, 2018c). From the data collection to data preparation, data analysis, and deployment of results and model, the issues of data analytics process integrate not only the scalability, or heterogeneity, or data quality …, but also how to protect and ensure the security of the data is also an important concern, especially when we consider that the analyzing of the data and the highlighting of the different correlations can disclose information that was meant to remain anonymous (Sedkaoui, 2018c):

- Heterogeneity: Transforming the several types of data, especially the unstructured data, to a format that allow its analysis is a big challenge in the analysis process. The heterogeneity and the big volume of data complicate the analysis of data. However, in their analysis process, the machine expects homogeneous data. In this case, data must be understood and structured to facilitate the analysis process.
- Scalability: In their attempt to analyze big data, new technologies are striving to satisfy a primordial property that is scalability. This means the ability of a system to improve its performance by increasing the size or number of its resources when faced with a larger load. So the challenges of big data analysis come from its large scale and also from the availability of mixed data based on different patterns or rules (heterogeneous mixture data) in the collected and stored data (heterogeneous mixture data issue) (Fujimaki and Morinaga, 2012).
- Timeliness: Time is important in big data context, because, companies need real-time information. It is about significant insights that data analysis can derive and advanced analytics methods and tools to better guide their decision-making process.
- Complexity: Decision-makers are increasingly confronted with problems of complexity of data. This complexity refers to the structure of data itself (widely unstructured). And any change in one element can upset the system and affect its behavior (Katal et al., 2013).
- Quality: More data doesn't mean always having the right data. So, the quality of data is an important element in the big data analysis process.
- Security: Protect the personal data collected from several devices and sources, is a big concern that the companies are facing in the big data context. They must secure personal information about their clients and ensure their use (analysis).

These challenges must be addressed carefully by companies because they present at the same time some technological and business issues. In this context, big data technologies are developed to respond to these challenges and go beyond traditional analysis tools. Because nowadays the structure of the data is different and produced strongly in an unstructured format and come not only from internal companies systems' but also from external varied sources.

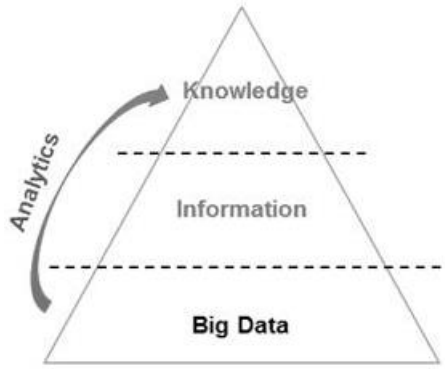

Figure 1. From big data to knowledge

Source: Sedkaoui, 2018a

Therefore, the data analysis process in the big amount of data age is a process characterized by its complexity. It focuses on the extraction of useful knowledge (see Figure 1) from a big volume of different data types (structured, semistructured, and unstructured data). This complex process goes through several phases that include data collection and preparation, data integration, aggregation and representation, query processing, modeling and analysis, interpretation and visualization.

Several analytics methods can be used to transform the available data to a form that facilitate the process of data analysis and allow valuable tendencies and correlations. There are many methods and algorithms that we can use to extract knowledge. Some of them can be summarized in Table 1. But, globally the objective of any data analysis process is to structure and organize data into significant format or taxonomy in order to obtain significant models able to enhance the decision-making process.

Table 1. Algorithms for data analysis

Objective	*Algorithm*	*Method*
X or Y or Z …	Supervised	Classification
How much? How many?	Supervised	Regression
The organization of the data?	Unsupervised	Clustering

Source: Sedkaoui, 2018b.

In fact, the data analytics process is more connected to the way the collected data needs to be treated and the consequential challenges must be addressed. Many researchers have shown that 'clustering techniques', allow extracting important insights and previously unknown understanding of the data in question.

3. Clustering algorithm

Machine learning is, of course, an important asset that allows us to deal with many society's issues and challenges. Among the various types of algorithms that we can find in this domain, we focus, in this section, on the application of data analysis that characterizes this domain: "clustering" (Sedkaoui, 2018b). It should be noticed that clustering analysis covers many topics and allow researchers to analyze the associated problems following their several perspectives. We can define cluster analysis as:

> *"a data mining process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters"* (Han et al., 2011).

To cluster is to group together similar objects according to certain criteria. The various clustering techniques all aim at distributing *n* individuals characterized by *p* variables $X_1, X_2,..., X_p$ into a certain number *K* that are as homogeneous as possible, each group being well differentiated from the others.

This method aims to identify a segmentation of the studied observation without a priori on the number of groups or what we call 'clusters' and interpret the created clusters. Clustering algorithms are most often used for exploratory data analysis and in several domains.

Cluster analysis involves the application of one or more clustering algorithms with the purpose of finding hidden patterns or groups in a dataset. Clustering algorithms can form classes or clusters so that cluster data has a higher

similarity. The similarity measure, by which the clusters are created, can be defined by a Euclidean distance, a probabilistic distance, or another metric.

Clustering analysis is used to identify inside a certain group of observations according to many properties. These proprieties can be age, level of education, sales, etc. Clustering aims to determine a segmentation of the studied population without a priori on the number of classes and to interpret a posteriori the groups thus created.

The application examples of this unsupervised learning algorithm are diverse and varied. In the world of business, we encounter this method through customer segmentation (see Figure 2), the subject of considerable importance in the marketing community.

Variable combinations are endless and make cluster analysis more or less selectively based on research requirements.

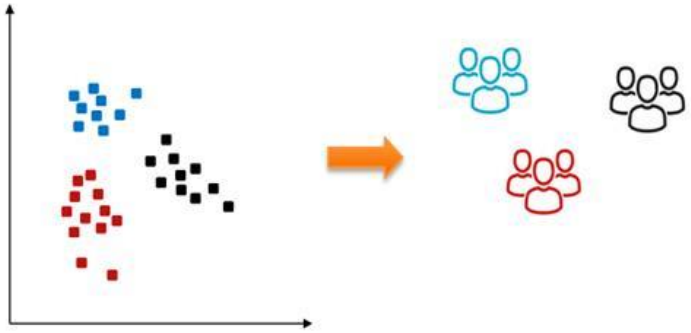

Figure 2. Clustering example

Source: Sedkaoui, 2018b

The aim is to facilitate data analysis and class them into the same cluster. So, the observations of the same cluster are similar, which means that classes are as distinct as possible.

Algorithms of clustering analysis examine a defined number of data properties and form groups of individuals or variables in order to structure a set of data. Clustering analysis is an unsupervised learning algorithm and is distinguished among other things by the clustering structure obtained (partition, hierarchy etc.).

The goal is to divide a set of objects, represented by inputs:

$$\{x_1, x_1, \ldots, x_1\}$$

into a set of disjoint classes or clusters:

$$\{\{x_{1,1}, x_{1,2}, \ldots, x_{1,n},\}, \{x_{2,1}, x_{2,2}, \ldots, x_{1,n}\}, \ldots\ldots\ldots, \{x_{n,1}, x_{n,2}, \ldots, x_{n,n}\}\}$$

That contains objects similar to each other in some sense.

This method aims to group the data into similar groups called clusters. In other word, this is to say that the objects of the same class must be "similar" and the objects of two different classes must be "distinct".

In this case, it should be noticed that it is important to identify a measure of similarity between two elements of the data (the distance). Each element can be defined by the values of its attributes, or what we call, from a mathematical point of view, 'a vector'.

$$U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u \end{bmatrix}$$

The purpose of this type of analysis is to segment the unstructured data. For this, algorithms are applied. The algorithms review data quantities, find structural similarities, and thus identify different clusters.

Clustering algorithms therefore strongly depend on how we define this notion of similarity, which is often specific to the application domain. The principle of the algorithm consists to assign classes according to (Sedkaoui, 2018a):

- The highest internal homogeneity (within each class);
- The highest external heterogeneity (among the different classes).

This similarity corresponds respectively to the internal variance (within the cluster) and the external variance (between clusters).

The key point so is the similarity criterion (distance function). However, to achieve this analysis process we must to try all possible combinations and choose the solution with the minimum intra-class distances, and the maximum inter-class distances. Many other clustering algorithms are developed for big data analysis (Sedkaoui, 2018b) to generate knowledge from data (see Table 2)

4. Clustering Algorithms for Big Data

In the big data context, there are two types of methods based on the number of computer nodes that have been used (see Figure 3):

i- Single-machine techniques and;
ii- Multi-machine techniques.

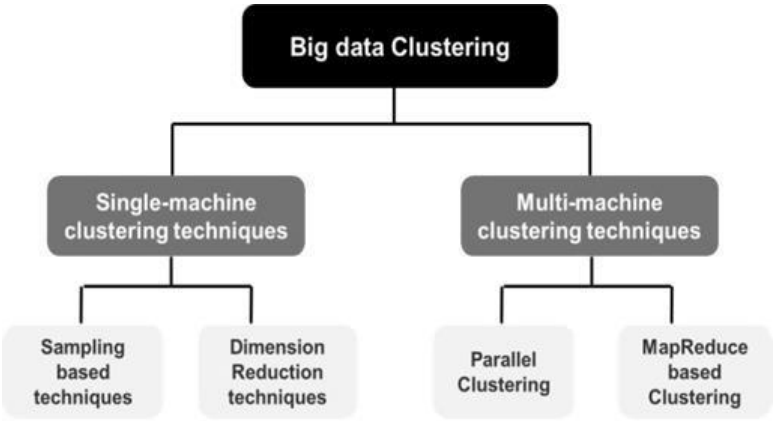

Figure 3. Clustering techniques in big data

Regarding the scalability nature and real-time analysis need to respond to the users, many machine clustering analysis have been developed and gain more importance. It should be noticed that the procedures of the cluster analysis contain the following steps:

- Select, or extract, the feature
- Select the cluster analysis
- Validation of the designed cluster analysis
- Interpret the results

These procedures reduce the size of the involved issues and can be applied in the extraction and the discovery of the hidden insights in the amount of available data. This algorithm makes it possible by dividing the data into correlative classes. Regarding data characteristics or the objective of the different clustering analysis, many algorithms have been created and developed. Basically, clustering analysis can be classified into the following types:

Table 2. Clustering categories

Category	*Characteristic*
Partition method	Create an initial partition of K classes, then iteration of a process that optimizes partitioning by moving objects from one class to another
Hierarchical method	Clusters are organized as a tree structure Choice of a distance criterion between clusters and a strategy for aggregation / division
Density-based method	Using density instead of distance Clusters are areas of space that have a high density of points
Grid-based method	Method based on the division of the space of the examples according to a grid
Model-based method	Clustering analysis assumes that the data is produced by a mixture of underlying probability distributions
Evolutionary method	Use genetic algorithm, particle swarm optimization, and other evolutionary approach for clustering task

In the big data age, many clustering algorithms can be used. But, the K-means algorithm is probably the best-known clustering technique. This algorithm splits the data into k separate clusters according to the distance with the centroid of the cluster. We can use also:

- Hierarchical method: which helps to develop a multilevel cluster hierarchy by creating a cluster tree
- The Gaussian mixing model: that models clusters as a mixture of multivariate normal density components
- Kernel K-means
- Spectral Clustering
- …

The feature that distinguishes each of these algorithms is the metrics for measuring similarity. But, basically, the different clustering methods seek to achieve these two objectives:

- identify the number of groups or clusters
- Find the structure of the groups or clusters.

But, many applications show that identifying this number before using clustering analysis is a very serious challenge. In this order, we can claim that two better apply the clustering algorithms we should first find the clustering tendency. This is to say, find how many cluster *K* are presented in the large volume of data.

5. Clustering Assessment Tendency

It is evident that if we know, in the beginning, the number of *K* clusters; we can apply any cluster techniques to analyze data. To define this number we have also, at our disposal, many cluster validity metrics. But what can we do when the number of clusters (*K)* is not predefined. In this situation, we must find a solution to answer the above question.

So, among the big data challenges, defining the number of clusters is also an important concern. Therefore, the Assessment of clustering tendency is an important topic and needs more attention in the clustering analysis process. So, if we can identify this number the process will be easier to achieve. This is possible, of course, by applying some techniques that experts have developed to respond to this concern.

Table 3. The various developed VAT

Algorithm	*Literature*
big-VAT	Bezdek and Hathaway, 2005
re-VAT	Huband et al., 2004
s-VAT	Hathaway et al., 2006
co-VAT	Bezdek et al, 2007
o-VAT	Pakhira, 2010
i-VAT	Wang et al., 2010; Havens and Bezdek, 2012

Many attempts have already been made to estimate the number of clusters present in a large volume of data. These methods include *"split-merge techniques"* and *"validity index based techniques"* mostly.

To face this situation, Bezdek and Hathaway have developed an important algorithm, called the VAT algorithm (Visual Assessment of Tendency) to display reordered dissimilarity data. The original VAT algorithm introduced by Bezdek and Hathaway in 2002 provides a useful visual display of well-separated cluster structure.

The VAT algorithms have been created to facilitate the determination and the identification of *K* cluster. In the literature, we can find a large variety of application of VAT (illustrated in Table 3) which have been proposed by experts and researchers.

These varieties of visual algorithms, developed in the context of clustering analysis, allow us to find out the approximate number of clusters. These algorithms must be applied before the application of the clustering analysis to facilitate tasks. The different VAT algorithms are developed based on the visual approach, i.e., VAT outputs are plotted as images on the output devices.

For example, in the hierarchical clustering analysis, we can use a SHADE visual technique. This technique is a close relative of the VAT algorithm.

In VAT we work with a pairwise distance matrix of the original object set:

$$\theta = \{\theta_1, \theta_2, \dots \theta_n\} \quad (1)$$

In the ijth element of the distance matrix pairwise similarities $S = [S_{ij}]$, then dissimilarities can be obtained by a simple transformation (Kendall and Gibbons, 1990), like:

$$d_{ij} = S_{max} - S_{ij} \quad (2)$$

Where, S_{max} denotes the largest similarity value.

We can illustrate a simple case about the VAT algorithm by appealing a graphical form shown as follow:

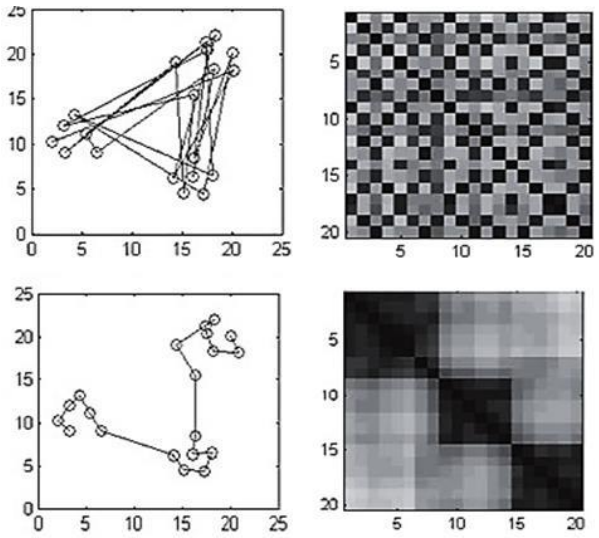

Figure 4. VAT Example

Source: Bezdek and Hathaway, 2002

When, the first graph presents a sample data displayed as a Graph and its dissimilarity image. The second one refers to the reordered graph and its dissimilarity image. From Figure 4, we can identify three clusters in the large volume of data which are represented by the three dark square blocks along the diagonal line.

6. Conclusion

Big data needs to consider complex relationships between samples, models and data sources. And, clustering analysis is one of the important methods that we can be used to analyze a large volume of the data and address big data challenges. To use the appropriate clustering method solely depends upon our requirement, the application involved and the other governing factors.

But, we need to determine the number of clusters present in a data set to better achieve the clustering analysis. Traditional known methods cannot allow us to define this number. Because of this researchers developed some techniques in order to solve this problem. These techniques can be used to detect automatically the number of clusters in a large amount of data. Some of these techniques rely on user-supplied information, while some others use cluster validity indices which are expensive with regard to computation time.

In this context, visual methods have been widely studied and used in data cluster analysis, and one tool for assessing cluster tendency is the Visual Assessment of Tendency (VAT) algorithm. This algorithm can be useful because it produces a visual aspect that helps to determine the clustering tendency in either relational or object data.

REFERENCES

Bezdek, J.C., Hathaway, R., Huband, J. (2007). Visual assessment of clustering tendency for rectangular dissimilarity matrices. *IEEE Transactions on Fuzzy Systems*, 15(5) 890–903

Bezdek, J. C., and Hathaway, R. J. (2005). bigVAT: visual assessment of cluster tendency for large data set, *in Pattern Recognition*, 38 (11), pp. 1875-1886

Bezdek, J.C., and Hathaway, R. J. (2002) . VAT: A tool for visual assessment of (cluster) tendency, *in Proc. Intl. Joint Conf. on Neural Networks.* Honohulu, HI, pp. 2225-2230.

Boyd, D., and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5): 662-679.

Ekbia, H., Mattioli, M., Kouper, I. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology* , 66(8), 1523-1545.

Fujimaki, R., and Morinaga, S. (2012). The Most Advanced Data Mining of the Big Data Era, *Advanced technologies to support big data processing*, 7 (2)

Han, J., Jian, P., and Micheline, K. (2011). *Data Mining: Concepts and Techniques. Burlington*, MA: Elsevier.

Hastie, T., James, G., Witten, D., and Tibshirani, R.(2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Hathaway, R., Bezdek, J. C., and Huband, J. M. (2006). Scalable Visual Assessment of Cluster Tendency, *in Pattern Recognition*, 39, pp. 1315-1324

Havens, T. C. and Bezdek, J. C. (2012). An efficient formulation of the improved visual assessment of cluster tendency (iVAT) algorithm, Knowledge and Data Engineering, *IEEE Transactions*, 24 (5), pp. 813–822

Huband, J. M., Bezdek, J. C., and Hathaway, R. (2004). Revised Visual assessment of (cluster) tendency (reVAT), *in Proc. Of NAFIPS*, pp. 101-104

Katal, A., Wazid, M., and Goudar, R.H. (2013). Big Data: Issues, Challenges, Tools and Good Practices, *IEEE Spectrum*, 404-409

Kendall, M., and Gibbons, J.D. (1990). *Rank Correlation Methods*. Oxford University Press, New York

Pakhira, M. K. (2010). Out-of-Core Assessment of Clustering Tendency for Large Data Sets," *in Proc. of the nd Int. Conf. on Advance Computing and Communications*, pp. 29-33

Sedkaoui, S. (2018a). *Data analytics and big data*, London: ISTE-Wiley.

Sedkaoui, S. (2018b). *Big Data Analytics for Entrepreneurial Success: Emerging Research and Opportunities*, New York: IGI Global.

Sedkaoui, S. (2018c). Statistical and Computational Needs for Big Data Challenges. In A. Al Mazari (Ed.), *Big Data Analytics in HIV/AIDS Research* (pp. 21-53). Hershey, PA: IGI Global. doi:10.4018/978-1-5225-3203-3.ch002

Sedkaoui, S., and Gottinger, H-W. (2017). The Internet, Data Analytics and Big Data, *In Internet Economics: Models, Mechanisms and Management* (pp. 92-105), Hans-Werner GOTTINGER: by eBook Bentham science.

Wang, L., Nguyen, T., Bezdek, J., Leckie, C., and Ramamohanarao, K. (2010). iVAT and aVAT: enhanced visual analysis for cluster tendency assessment, *in Proc. PAKDD, Hyderabad*, India, Jun. 2010.