Using Big Data Classification and Mining for the Decision-making 2.0 Process

Rhizlane Seltani[1,2]
[1,2]Information Technology and Modeling Systems Research Unit, LIROSA Laboratory
Faculty of Science, Abdelmalek Essaadi University
Tetuan, Morocco
sel.rhizlane@gmail.com

Noura Aknin[1,2]
[1,2]Information Technology and Modeling Systems Research Unit, LIROSA Laboratory
Faculty of Science, Abdelmalek Essaadi University
Tetuan, Morocco
aknin@ieee.org

Souad Amjad[1,2]
[1,2]Information Technology and Modeling Systems Research Unit, LIROSA Laboratory
Faculty of Science, Abdelmalek Essaadi University
Tetuan, Morocco
amjad_souad@uae.ma

Kamal Eddine El Kadiri[2]
[2]Computer Science, Operational Research and Applied Statistics Laboratory
Faculty of Science, Abdelmalek Essaadi University
Tetuan, Morocco
elkadiri@uae.ma

***Abstract*—Web 2.0 is a revolution that has affected all areas, especially those of the new technology. Several new concepts have emerged, and a large number of innovative applications continue to come out every day. However, the social networking remains the racehorse of web 2.0, giving the user at the same time, a space for communication and for information sharing, which generates too much data, variable and characterized by a great creation speed. So, we can call them big data, and consider them a very rich and interesting basis for decision-making.**

Big Data is a type of data which are characterized by the veracity, important volumes, and increasing variety and velocity, which makes their treatment and their processing by traditional database management tools a very difficult task. To overcome this problem, we opt for the big data classification process.

In this paper, we make a study of some big data classification methods, which are the most significant to be used to classify big data dedicated to decision-making, we detect their points of strength and weakness. Then we propose a framework summarizing the process of the formulation of the decision from the web 2.0 content, based on the big data classification, and we specify the criteria to be taken into account when choosing the big data classification methods intended for the decision-making.

Keywords—Web 2.0; Big Data; Decision-making; Data Classification

I. INTRODUCTION

The large variety of applications that appeared after the emergence of the web 2.0, produce a huge mass of various and diverse data. This wealth of information is a very important resource that we want to exploit to enrich our decision-making systems, to generate more meaningful and relevant decisions.

To classify and process data, various algorithms and techniques can be used. These methods differ depending on data types. In the case of big data, to retrieve information, there are various analysis techniques with different orientations and results, such as Representation-learning methods based geometric information, Stream Classification Algorithms, Associative Classifiers, etc.

In this paper, we discuss some methods that we can use to classify big data in order to elaborate decisions, report the strengths and the weaknesses. And therefore, present our global framework of decision-making 2.0 based on big data classification by describing the key pillars to be considered, to lead well the classification process for the purpose of decision-making.

II. WEB 2.0

A. Definition

The web 2.0 is a combination of technologies, business plans and social skills, which allow users to create web content, and to be more involved in the process of the management of this content. It has brought many creative concepts and techniques that did not exist before and which made the electronic life simpler and more enjoyable [1][2]. With the web 2.0, a new era of web use is born. Several applications have been developed and which have enriched our lives by allowing more of interactivity and collaboration, such as blogs and social networks [3].

B. *Architecture and Principals*

Web 2.0 is based on a varied and robust architecture, founded on the introduction of new principles such as collaboration and interactivity, and the use of new applications like web interface design techniques, those of content syndication, XHTML, URL, etc [4].

There are several emerging principles with the appearance of web 2.0, the most notable:

- Collaboration: This is an important aspect which describes when a user has the opportunity to contribute in the creation of the web content by creating its own content.
- Interactivity: one of the introduced principles by the web 2.0, interactivity is reflected by the interaction of the user with the web content and with other users.

These two principles constitute new trends that have changed our lives and our way of working, they are the basis of social networks, blogs, wikis, etc.

III. BIG DATA

A. *Definition*

The term big data refers to data sets exchanged by connected objects in the web, and whose volumes are important and the variety and the velocity are increased [5]. It is a compilation of data sets which are characterized by complexity and large volume, so their management and processing constitute a difficult task if we use traditional database management tools [6].

B. *Characteristics*

Compared to other types of data, big data are different and have some specifications. These differences concern several facets as the data format, their volume, the time required for their creation, and their nature.

The principal features are: Data volume, data velocity, data variety, and data veracity. We can consider these elements as the characterizing pillars of big data (Fig. 1.), and which make their processing and their analysis a special challenge.

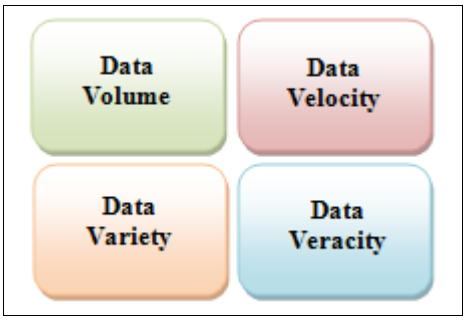

Fig. 1. Characterizing Pillars of Big Data

- **Data Volume**: refers to a very important quantity of generated information. Data is considered as big data if their size is very large, so we cannot control them to make analysis easily.
- **Data Variety**: This makes analyzing this type of data a very difficult mission. We have more different data presentation formats: text, audio, image, etc.
- **Data Velocity**: It refers to the speed of creation and generation of data, which have been increased with the different new web applications.
- **Data Veracity:** Data veracity refers to the anomalies in data. Veracity in data analysis constitutes the biggest challenge to overcome, because, veracity of data sources can largely affect the precision of analyzes.

IV. BIG DATA CLASSIFICATION FOR DECISION-MAKING

A. *Clustering*

Clustering (also called Cluster Analysis), is a task of data mining, which means the mission of assembling a set of objects, by the way that, objects which belong to the same group have more similarities than with those belonging to others groups. A group is called a cluster. The clustering was used for the first time in the classification tasks by Cattell in 1943 for personality psychology classification [7]. Many clustering algorithms exist. Making the choice about which algorithm we must use, depends on the used cluster models [8]. Among the most distinctive cluster models, we find: Centroid models, Distribution models, Group models, and Connectivity models.

In addition to its important role in the classification task, clustering has several advantages, such as the definition of information relating to the data, which were not revealed before, as associations, so we can look for new patterns. Also, clustering provides a logical structure which makes results read and interpreted easily. But it is not the case, if we opt for a large scale of clusters, because there are no definitive methods to determine precisely the suitable number of clusters.

B. *Decision Trees*

The decision tree is a technique which we can use for classification tasks, by creating a model to predict the output value based on a number of input values [9] [10]. To use decision trees for classification, we construct trees starting by the root of the tree, and subsequently, proceeding down to its leaves.

A classification rule is developed based on example objects, which are known by their values of a collection of attributes. Then, the decision tree is expressed in function of the same attributes [11]. Decision trees constitute a good way to well represent decisions. An example of a decision tree form is shown in the Fig. 2.

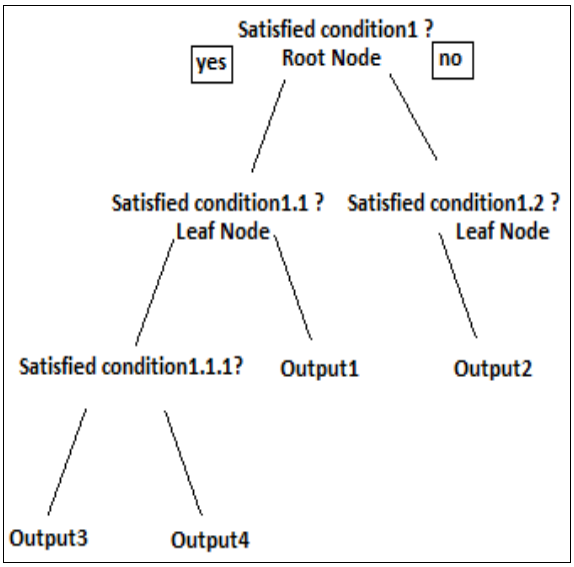

Fig. 2. A General Form of a Decision Tree

The decision trees are characterized by the robustness and the simplicity of understanding and interpreting. What is important about decision trees is that we can treat categorical and numerical data. On the other hand, decision trees are instable, since a miniature change in the input data can affect the entire tree, by causing large changes in it.

C. Support Vector Machines

Support vector machines, more usually SVMs, were introduced the first time for binary classification. They refer to a collection of methods used for regression and classification, to analyze data in order to verify to which category an element belongs [12]. They can be used in several ways depending on the nature of their application, such as, text categorization, recognition of images, hand-writing code, bioinformatics, etc.

Some of the advantages of using SVM algorithms are: the robustness, the ability to learn well using a few parameters, and the computational efficiency. On the other hand, apply SVM can at times require taking into consideration many aspects of learning methods [13], SVM is oriented to be applicable directly in the case of two-class tasks. For that reason, when we deal with a multi-class task, we must use algorithms that can reduce it to a set of binary problems, or take account of all the classes at once by giving one formulation of optimization for all the data. Different methods of treating multi-class support vector machines continue to emerge [14].

D. Associative Classification

Associative classification refers to a classification which is based on the use of association rules, by combining both classification and mining of associations [15] [16]. Compared to other approaches, it is considered a highly accurate and competitive method, and can be applied in different ways [17] [18] [19] [20]. We can define three types of associative classification systems:

- Classification by Emerging Patterns: based on emerging patterns from a sample, which means event associations whose supports vary, depending on the dataset [21].
- Classification based on High-Order Pattern: is a classification system, which uses the algorithm of high-order pattern discovery, which detects considerable connection or association patterns using residual analysis in statistics [22].
- Associative Classifiers based on the Apriori Algorithm: the Apriori Algorithm is an algorithm which proceeds by determining the prevalent items in the database. So, we can define association rules to wrap up trends in the database, many applications in various domains were done using this technique, such as market basket analysis [23].

Associative classification provides a high accuracy and it is easy to understand. However, it presents some challenges, like the lack of obvious criteria to classify objects. Since it is based on a large number of rules, the process of its elaboration is a time-consuming task, and it becomes a difficult task to select the suitable ones to develop the classifier.

V. Big Data Classification as a Basis of Decision-making 2.0

A. The Data Generation Process

Web 2.0 is a very important source of information. The user interacts continuously with the web content through collaborative applications, such as blogs, social networks, etc. With the increase of the number of actors on the web, the rate of information circulating on its channels increases. This large data flow generates the phenomenon of big data. Hence, web 2.0 is a rich platform of information, which can be treated to generate significant data. The user is primarily a passive actor, becomes in an instant an active actor, by transmitting opinions, which we propose to treat to ensure the mission of decision-making. These opinions can take, for example, the form of:

- A solution to a particular problem: a problem can be solved quickly and efficiently if the process of the generation of the solution is collaborative. So the reviews, including those of experts, about an issue may be of great use to make decisions to solve a given problem.
- A feedback to a given subject: any feedback contains in itself a notice that we can use to extract useful information which enriches the process of the decision making.
- A proposal for improvement: in any field, application, or system, we always look for ways of improvement, especially in the case of business. Opinions of clients and in particular those which are

the most affected by the service, constitute a very important resource of inspiration to make the right decision of improvement.

- A complaint about a process, a product, a service: as with proposals for improvement, complaints also lead to the generation of significant decisions about a product, a process, a service, etc.

B. Decision-Making 2.0 Based Big Data Classification Model

To exploit the generated data on the web 2.0, it is necessary to isolate the significant information. Circulating data through the web 2.0 applications such as social networks have the characteristics that make them a part of what is called big data. To process them, we proposed to adopt a classification process.

When we want to treat data based on the web 2.0 content, in order to make decisions. A simple comment or tweet can generate a large data stream, through feedbacks of users. Taking account of these data in decision-making is very important to harness the collective intelligence.

After a preliminary process of data streams, to centralize those that meet our study needs, comes the classification phase to derive classified data according to specific parameters that depend on the issue in question. Finally, we get the basis of decision-making. The framework which presents the general process starting with the creation of the data on the web and ending with the decision-making is represented in the Fig. 3.

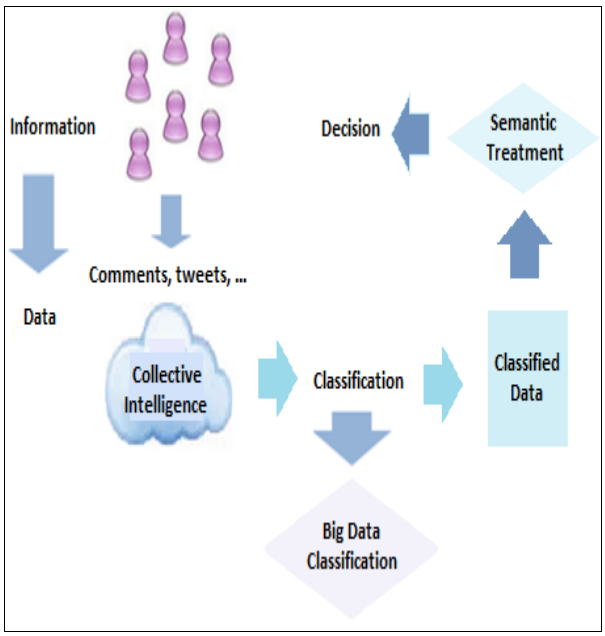

Fig. 3. Process of the Generation of the Decision 2.0 Based on the Big Data Classification

In the decision-making 2.0 process, the classification serves as a passage from the raw data to the classified ones, which will be used later to generate decisions. Data which circulate across the web, especially in social networks, blogs, etc, are difficult to track and manage. So to overcome this problem, our classification process should follow some specifications to properly carry out this mission.

Taking into consideration our aim, which is decision-making based on the content reflected by the comments and the feedbacks of users, and to provide relevant decision, which must be generated based on meaningful data, our classification process must be efficient and suits our purpose.

As already mentioned, the classification methods have drawbacks as advantages. That is why, we opt for a combination, to elaborate a multiple classification model to exploit the strengths of the cited methods, taking into account different parameters, as shown in the Fig. 4.

- **Accuracy:** the classification process must guarantee high accuracy, to ensure the relevance of our decisions, which is a very important factor for the evaluation of the quality of the decision.
- **Facility of understanding:** it is essential that classification must be a process that provides results which are easy to understand. It means also, that results must be interpreted without difficulties.
- **Flexibility:** flexibility is represented by the fact that the classification can take into consideration categorical data, and not just the numerical ones, for more significant and common decisions.

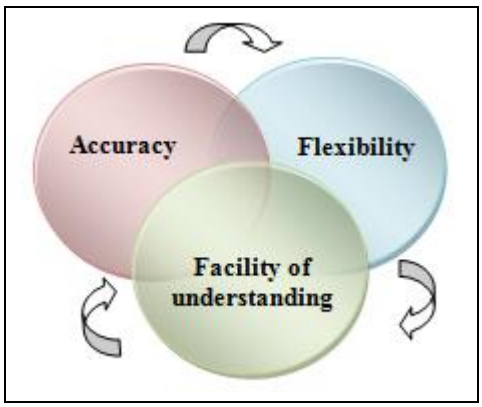

Fig. 4. Pillars of Big Data Classification for a Decision-making 2.0 Process Model

VI. Conclusion

In this paper, we gave a vision on the results of a developed study of the big data classification tools, we presented a summary of the results concerning the techniques that we can use to treat data coming from web 2.0, to ensure the decision-making mission. Then, we presented a general framework of the entire process and mentioned the criteria to take into consideration when choosing the classification method.

To exploit the strengths of the cited methods, we opt for a combination, to develop a multiple classification model, so that we can ensure three pillars of big data classification for a decision-making 2.0 process, which are accuracy, facility of understanding and flexibility.

Acknowledgment

The authors of this paper would like to thank our Research Team, Information Technology and Modeling Systems Research Unit, and more generally, the Computer Science, Operational Research and Applied Statistics Laboratory, from the Faculty of Science, Abdelmalek Essaadi University of Tetuan, Morocco, for their great support.

References

[1] T. O'reilly, "What is Web 2.0: Design patterns and business models for the next generation of software." Communications & strategies, (1), 17, 2007.

[2] T. O'Reilly, and J. Musser, Web 2.0 principles and best practices. O'Reilly Radar, 2006.

[3] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks." In Link mining: models, algorithms, and applications, Springer New York 2010, pp. 337-357.

[4] T. O'Reilly, What is web 2.0. O'Reilly Media, Inc, 2009.

[5] P. Zikopoulos, and C. Eaton, Understanding big data. Analytics for enterprise class hadoop and streaming data, 2012.

[6] E. Letouzé, "Big data for development: challenges & opportunities". UN Global Pulse, 47, 2012.

[7] R. B. Cattell, "The description of personality: basic traits resolved into clusters." Journal of Abnormal and Social Psychology 38: 476–506, 1943.

[8] V. Estivill-Castro, "Why so many clustering algorithms — a position paper." ACM SIGKDD Explorations Newsletter 4 (1): 65–75, 2002.

[9] L. Rokach, Data mining with decision trees: theory and applications. World Scientific Pub Co Inc. ISBN 978-9812771711, 2008.

[10] S. B. Kotsiantis, "Decision trees: a recent overview." Artificial Intelligence Review, 39(4), 261-283, 2013.

[11] J. R. Quinlan, "Induction of decision trees." Machine learning, 1(1), 81-106,1986.

[12] V. N. Vapnik, The nature of statistical learning. Springer-Verlag New York, 1995.

[13] I. Steinwart, and A. Christmann, Support vector machines. Springer Science & Business Media, 2008.

[14] C. W. Hsu, and C. J. Lin, "A comparison of methods for multiclass support vector machines." Neural Networks, IEEE Transactions on, 13(2), 415-425, 2002.

[15] Y. Wang, and A. K. C. Wong, "From association to classification: Inference using weight of evidence." IEEE Trans. On Knowledge and Data Engineering, 15(3):764–767, 2003.

[16] X. Yin, and J. Han, "CPAR: Classification based on predictive association rules." In Proceedings 2003 SIAM International Conference on Data Mining(SDM'03), San Francisco, CA, May 2003, pp. 331– 335.

[17] G. Dong, X. Zhang, L. Wong, and J. Li, "CAEP:classification by aggregating emerging patterns." In Proceedings of The Second International Conference on Discovery Science (DS'99), pp. 43–55, Japan, December 1999.

[18] J. Li, G. Dong, K. Ramamohanarao, and L. Wong, "DeEPS: a new instance-based lazy discovery and classification system." Machine Learning, 54(2):99–124, 2004.

[19] W. Li, J. Han, and J. Pei, "CMAR: accurate and efficient classification based on multiple class-association rules." In Proceedings of The 2001 IEEE International Conference on Data Mining (ICDM'01), pp. 369–376, San Jose, CA, November 2001.

[20] B.L.W.H.Y. Ma, "Integrating classification and association rule mining." In Proceedings of the Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 80–86, New York, NY, August 1998.

[21] G. Dong and J. Li. "Efficient mining of emerging patterns: discovering trends and differences." In S. Chaudhui and D. Madigan, editors, Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 43–52. ACM Press, San Diego, CA, 1999.

[22] Y. Wang, High-order pattern discovery and analysis of discrete-valued data sets. PhD thesis, University of Waterloo, Waterloo, Ontario, Canada, 1997.

[23] R. Agrawal, and R. Srikant, "Fast algorithms for mining association rules." In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499, 1994.