Data mining for road accident analysis in a big data context

Fatima Zahra El Mazouri[1], Mohammed Chaouki Abounaima[2], Said Najah[3], Khalid Zenkouar[4]
{Fatimazahra.elmazouri@usmba.ac.ma[1], medchaouki.abounaima@usmba.ac.ma[2],
Said.najah@usmba.ac.ma[3], khalid.zenkouar@usmba.ac.ma[4]}

Laboratory of Intelligent Systems and Applications, Faculty of Sciences and Technologies, Sidi Mohammed Ben Abdellah University, Fez, Morocco [1, 2, 3, 4]

Abstract. Data Mining techniques and extracting association rules from a big dataset play an interesting role in knowledge discovery. Therefore, the decision makers encounter a huge number of resulting association rules that can make them unable to choose and decide rationally between these different extracted rules, also the time of the generation of these association rules brings a new challenge, we propose to overcome these challenges a learning model based on FP-growth algorithm using Apache Spark framework, in order to analyze data and extract interesting association rules by taking into account some quality measures. Experimental results on road accident data in France show that the proposed approach can provide useful information that could help the decision makers to choose the appropriate strategies in the perspective of improving road safety.

Keywords: Data Mining, Association Rules, FP-growth, Big Data, Apache Spark, Road Accident.

1 Introduction

This work is considered as a continuity of work [1] in which we have combined data mining and multicriteria decision analysis approach, in this work we propose an approach based only on data mining with a refinement step in order to extract the most relevant association rules.

In the last years, a big interest is addressed to analyzing data, in order to keep significant information and extract useful knowledge for further use by the decision maker. Data mining [2] is considered as a very important step of Knowledge Discovery in Database process [3], data mining is a set of algorithms and methods for exploration and analysis large databases in order to extract rules, associations, unknown trends, etc. [4]. Searching for relations and correlations between objects in a large database is a very powerful technique of data mining for discovering relevant information [5].

Extracting association rules from transactional databases have received intensive research since its introduction by Rakesh Agrawal [6], it has been successfully used in various domains and applications such as basket analysis [7] bioinformatics [8] chemo-informatics [9] biomedical [10] image [11].

Since the datasets are extremely large, parallel algorithms are recommended, for these reasons we chose apache spark [12] as a powerful framework to process and analyze big datasets by exploiting machine learning algorithms like FP-growth (Frequent Pattern growth), this algorithm proposed by Han et al. [13] is a very efficient technique thus provides a solution to the problem of searching for frequent patterns and association rules in a large transactional database. In addition, it generates a big number of association rules when the transactional database is large. Therefore, it is necessary to help the decision maker to select the most interesting rules among a large number of extracted rules. We can exceed this difficulty by implementing a post-processing stage of extracted rules, which consists to use a refinement process. This process that we propose in this work is based on 2 steps of refinement, in the first step, we retain that the rules of associations which have the best support and confidence thresholds, in the second step, we retain this time the best association rules, result of step 1, which have the best lift threshold.

Many statistics and works [14,15] show the gravity of road traffic accidents and remain worrying for everyone, and we are seriously trying to find effective solutions to fight against these losses of human lives. In this context, our study provides help for the stakeholders and decision-makers and finds correlations between human damage and the various conditions that characterize all recorded accidents. The aim of this work is to treat the case of road accidents in France by using the BAAC (Accident Analysis Bulletin Corporal) database [16-17], this database made by the French ministry of interior and the ministry of transport, our purpose is to use BAAC database in order to inform decision-makers of the most common conditions at the origin of accidents. To do this, we propose to extract the association rules by using the FP-growth algorithm in a big data environment using Apache Spark. Given a large

ICCWCS 2019, April 24-25, Kenitra, Morocco

DOI 10.4108/eai.24-4-2019.2284124

number of the generated association rules, therefore, their exploitation becomes more and more difficult, which leads us to use a refinement process based on support, confidence and lift measures.

The rest of this paper is organized as follows: the two first sections "association rules" and "Apache spark" are designed to give an overview of the association rule techniques and apache spark framework. "Proposed approach" section describes the proposed methodology. The results and discussion are presented in the "Results and evaluation" section. In the last section, we concluded by summarizing the work and giving some perspective in the "conclusion" section.

2 Association rules

2.1 Preliminary

Consider a transaction database D, in which each row contains a set of items (itemset/pattern) I with its identifier (Tid), The Items: A, B, C … each item can design a product (in other context books, cars, projects.)

A pattern or itemset is a set of items: C, AB, BCD...

The generation of association rules from a transactional database consists to extract a rule in the form A → B (A, B ⊂). Every rule is composed of two different sets of items, also known as itemset, A and B, where A is called antecedent or left-hand-side (LHS) and B consequent or right-hand-side (RHS).

An example of a rule A → B mean IF A THEN B: which represent a relation between A and B.

- The support of rule: support(A→B) = support(A∪B), means the support of the union of the itemsets in A and B. with:

$$\text{Support(A)}=\frac{|\{t \in T; A \subseteq t\}|}{|T|}, \quad (1)$$

the support of A with respect to T (with T is the table of transactions) is defined as the proportion of transactions t in the dataset which contains the itemset A.

- The confidence determines how frequently items in B appear in the transactions that contain A, ranges from 0 to 1.

$$\text{Confidence (A}\rightarrow\text{B)} = \frac{\text{support (A}\cup\text{B)}}{\text{support(A)}}. \quad (2)$$

- The lift measures how far from independence are A and B. It ranges within [0, +∞]:

$$\text{Lift(A}\rightarrow\text{B)} = \frac{\text{support (A}\cup\text{B)}}{\text{support(A)}*\text{support(B)}}. \quad (3)$$

2.2 Association rules Algorithms

Many algorithms [**18**] have been designed for generating association rules, we can cite that APRIORI and FP-Growth, are the two most popular for finding association rules in data sets, Moreover, they prove their performance on a transactional database where each transaction design a pattern (an itemset). In [19] a study shows that FP-growth provides much more consistent and quicker performance then APRIORI, because that FP- growth uses divide and conquer strategy and it needs only two passes over the datasets so it is much more scalable.

2.3 FP-Growth algorithm

FP-growth is a data mining algorithm proposed by Han and al. [13], based on a data structure known as the FP-tree (It represents the frequent pattern tree), FP-growth has a major step for the extraction of frequent itemset from this structure: it compresses the frequent itemset represented in the database using an FP-Tree (frequent-pattern tree) whose branches contain the possible associations of the items. Each association can be divided into fragments (pattern fragment) which constitute the frequent itemset. The FP-Growth method transforms the problem of finding the longest frequent itemset by searching for the smaller one and its concatenation with the corresponding suffix (the last frequent item in the branch leading to the item in question) which leads to reduce the cost of research.

FP-growth algorithm:

- Use a compressed presentation of the database using an FP-tree.
- Once an FP-tree has been constructed, it uses a recursive divide-and-conquer approach to mine the frequent itemsets.

Building the FP-Tree

1. Scan data to determine the support count of each item.
   Infrequent items are discarded, while the frequent items are sorted in decreasing support counts.
2. Make a second pass over the data to construct the FP-Tree.
   As the transactions are read, before being processed, their items are sorted according to the above order.

3 Apache Spark

Apache Spark is an open-source cluster computing framework for real-time processing [12] this platform offers a variety of machine learning and data mining programs; these programs are very efficient and able to process and to analyze a large amount of data. Apache Spark has an RDD (Resilient Distributed Dataset) an immutable distributed collection of data, partitioned across nodes in a cluster that can be operated in parallel. Apache spark provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs.

Spark runs in master-slave mode, that is, a master and one or more workers. There is a driver that talks to a single coordinator named master that manages workers in which executors run. More clearly, when you start treatment on the Spark framework, you go through the Driver which is in a way the master, itself communicates with the cluster manager, the latter manages the resources of the workers (the workers execute the actual processing) (see Figure 1).

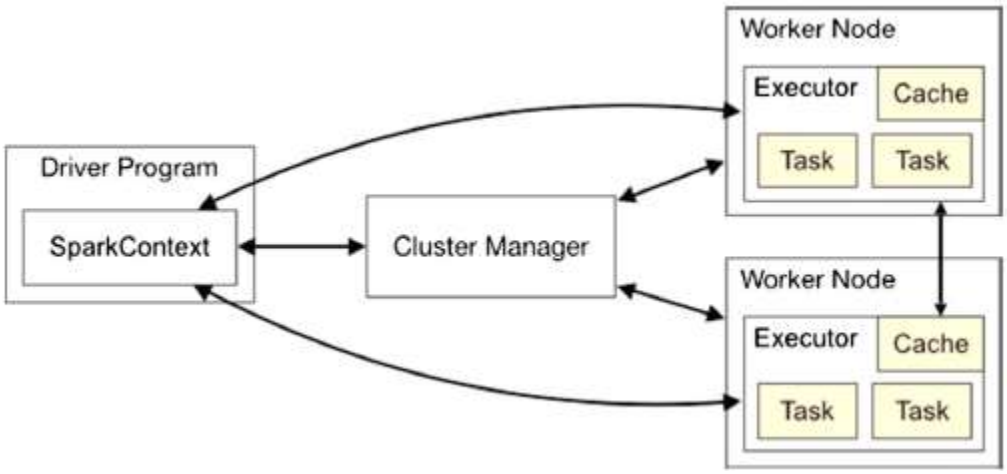


Fig. 1. Spark architecture.

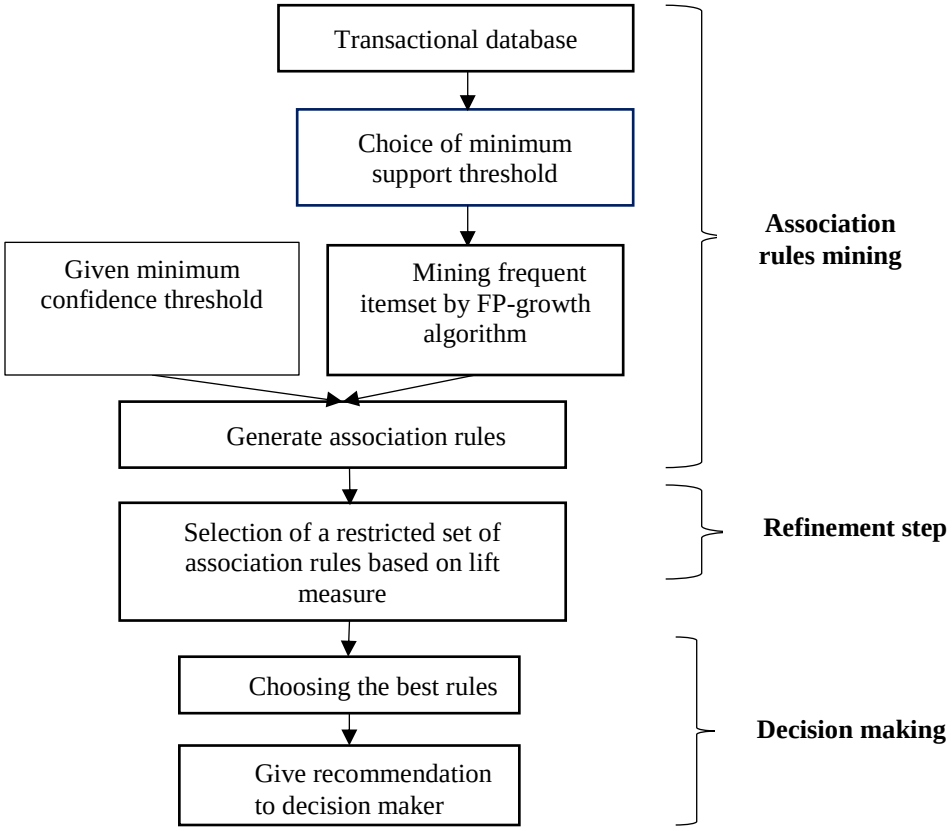
4 Proposed approach

In data mining, association rules algorithms produce a large number of rules that make the decision about the most interesting and useful association rules very difficult for the decision maker. To deal with this problem the integration of an evaluation of those association rules using some quality measures like support, confidence and lift can minimize the number of the generated association rules.

The purpose of our proposed approach is to extract knowledge about the most relevant conditions that cause the majority of road accident in France in the year 2017. According to the report published by ONISR (National Inter-Ministerial Observatory of Road Safety) on accidents occurred on the roads of France in 2017 [20], the cost of the corporeal accidents in metropolitan France would be 39.7 billion euros (B€) distributed as follows:

- 11.3 B€ in mortality;
- 23.1 B€ for hospitalizations;
- 4.0 B€ n for light victims;
- 1.2 B€ for the material damage of these bodily injuries.

The obtained association rule by the application of our approach is very useful to adopt for the right strategies to improve road safety in France.



The approach is summarized in 5 steps, the flowchart given in Figure 2 shows the full process.

Fig. 2. Flowchart of the proposed approach.

This proposed approach is described in the following steps:

Step1: collect data

We propose in this work, to use the databases collected on all the accidents that were recorded during the year 2017. The BAAC databases are available on the net [16-17], is composed of the 4 files in CSV format:

- Characteristics: This table describes the general circumstances of the accident. There are 16 attributes that are retained with 60 702 records.
- Places: This table describes the location where the accident occurred. There are also 18 attributes that are retained with 60 702 places described.
- Vehicles: This table describes the vehicles involved in the accident. There are also 9 attributes that are used for this table and we have 103 547 vehicles recorded.
- Users: This table gives all the information on the users involved in the accidents, a user can be a driver or a pedestrian. This table contains 12 attributes with 136 022 users registered.

Our first task was to create a database under the ORACLE DBMS for better exploitation of the data of 4 files (tables), we proceeded to the joining of the 4 tables to have a global vision on all the accidents. Finally, we obtained a transactions database with 135696 rows (accidents) and 31 attributes. Figure 3 shows an excerpt of the data.

	A	B	C	D	E	F
1	MONTH	HOUR	Light	Urbain	Intersection	Atmospheric_condition
2	January	Evening	Night-with-public-lighting-lights	In-urban	Out-of-intersection	Normal-ATM
3	January	Evening	Night-with-public-lighting-lights	In-urban	Out-of-intersection	Normal-ATM
4	January	Evening	Night-with-public-lighting-lights	In-urban	Out-of-intersection	Normal-ATM
5	February	Evening	Fullday	In-urban	Intersection-in-T	Normal-ATM
6	February	Evening	Fullday	In-urban	Intersection-in-T	Normal-ATM
7	March	Morning	Fullday	In-urban	Other_intersection	Normal-ATM

Fig. 3. An extract of the transactional database.

Step 2: Application of FP-growth algorithm on Apache Spark environment for association rules extraction

To have an appropriate model for extracting association rules from our database, we used the FP-growth algorithm on Apache spark environment. First, we get the frequent itemset by using minimum threshold support, then, the generation of the association rules based on the obtained frequent itemset and the minimum confidence threshold. In this study, we did the experiment with several thresholds (50%, 60%, 70%, 90%, and 100%) as shown in Table 1.

Step 3: Step of refinement

In this step, we try to choose the most relevant rule by crushing the thresholds of support and confidence until we obtain a set of minimal rules, and then, we get the most appropriate rule based on the rule with the highest lift.

Step 4: Recommendation of the final solution

At the last step, according to the classification of the association rules obtained by the application of the previous steps, we recommend to the decision-maker the most relevant associations in order to define, for example, the security measures to be considered in the short-term, median-term or in the long-term.

5 Result and discussion

At this stage, we present the analysis result of the treatment performed for association rules extraction, and we end by a refinement process to select the most interesting rule among the set of the obtained association rules. The experiments of our proposed approach are done by using road accident data in France in the year of 2017. The system was built on Spark by using single nodes.

5.1 Building the transactions table T

The obtained table of transactions T by merging the 4 tables (Characteristics - Places - Vehicles - Users) from the BAAC database, with an SQL join, as shown in Figure 3. This table contains 135 696 transactions, and each transaction is described in maximum by 220 items. Seen the size of the table of transactions, we present only an excerpt in Figure 3.

5.2 Application of FP-growth on apache spark environment

In this step, we chose the FP-growth algorithm in a big data environment using Apache Spark in order to generate association rules. Then, we look for the minimum support threshold and minimum confidence threshold giving the minimum number of association rules. The choice of thresholds minconf = 0.8 and minsup = 0.6 leads to the extraction of a minimum of association rules, a total of 15 association rules, which seems a more logical choice of the thresholds, see Table 1. But these 15 rules are extracted only at the base of the two measures support and confidence thresholds.

Table 1. Experimentation of several thresholds of confidence and support.

Support thresholds	Confidence Threshold	Number of extracted rules
0.5	0.5	116
0.6	0.5	19
0.7	0.5	2
0.8	0.5	0
0.9	0.5	0
0.5	0.6	116
0.6	0.6	19
0.7	0.6	2
0.8	0.6	0
0.9	0.6	0
0.5	0.7	94
0.6	0.7	19
0.7	0.7	2
0.8	0.7	0
0.9	0.7	0
0.5	0.8	60
0.6	**0.8**	**15**
0.7	0.8	2
0.8	0.8	0
0.9	0.8	0
0.5	0.9	16
0.6	0.9	4
0.7	0.9	2
0.8	0.9	0
0.9	0.9	0
0.5	1	0
0.6	1	0
0.7	1	0
0.8	1	0
0.9	1	0

Table 2 gives the 15 rules resulting from this step of association rules extraction with a minimum confidence threshold equal to 0.6 and a minimum support threshold equal to 0.8.

Table 2. The obtained association rules.

	LHS	RHS	Support	confidence
R1	{Drawing_in_plan=Straight-part}	{declivity_of_the_road=Dish}	0.606529	0.852739
R2	{Drawing_in_plan=Straight-part}	{Accident_Situation=On-the-pavement}	0.626383	0.880652
R3	{State_of_surface=Normal-surface}	{declivity_of_the_road=Dish}	0.636648	0.81309
R4	{declivity_of_the_road=Dish}	{State_of_surface=Normal-surface}	0.636648	0.81042
R5	{State_of_surface=Normal-surface}	{Accident_Situation=On-the-pavement}	0.636759	0.813231
R6	{State_of_surface=Normal-surface}	{Atmospheric_condition=Normal-ATM}	0.732776	0.935858
R7	{Atmospheric_condition=Normal-ATM}	{State_of_surface=Normal-surface}	0.732776	0.90892
R8	{declivity_of_the_road=Dish}	{Accident_Situation=On-the-pavement}	0.644666	0.820627

R9	{Accident_Situation=On-the-pavement}	{declivity_of_the_road=Dish}	0.644666	0.801176
R10	{declivity_of_the_road=Dish}	{Atmospheric_condition=Normal-ATM}	0.637931	0.812053
R11	{Accident_Situation=On-the-pavement}	{Atmospheric_condition=Normal-ATM}	0.649854	0.807624
R12	{Atmospheric_condition=Normal-ATM}	{Accident_Situation=On-the-pavement}	0.649854	0.806066
R13	{State_of_surface=Normal-surface,Accident_Situation=On-the-pavement}	{Atmospheric_condition=Normal-ATM}	0.600921	0.943719
R14	{Atmospheric_condition=Normal-ATM,State_of_surface=Normal-surface}	{Accident_Situation=On-the-pavement}	0.600921	0.820062
R15	{Atmospheric_condition=Normal-ATM,Accident_Situation=On-the-pavement}	{State_of_surface=Normal-surface}	0.600921	0.924701

5.3 Refinement Step

We rank the rules obtained in Table 2 using the lift measure to find the best association (Table 3).

Table 3. Association rules with lift measure.

Rules	Lift
R1	1.085493097
R2	1.094453198
R3	1.03502142
R4	1.03502142
R5	1.010664323
R6	**1.160819038**
R7	**1.160819038**
R8	1.019855463
R9	1.019855463
R10	1.007253184
R11	1.00175951
R12	1.00175951
R13	**1.170568809**
R14	1.019152938
R15	**1.180974197**

The first refinement step gave us 15 rules (Table 2) based on the 2 criteria: support threshold and confidence threshold. After we keep only 4 rules
R6:{State_of_surface=Normal-surface} → {Atmospheric_condition=Normal-ATM},
R7:{Atmospheric_condition=Normal-ATM} → {State_of_surface=Normal-surface},

R13:{State_of_surface=Normal-surface,Accident_Situation=On-the-pavement}→{Atmospheric-condition= Normal-ATM},
R15:{Atmospheric_condition=Normal-ATM,Accident_Situation=On-the-pavement} → {State_of_surface=Normal-surface},
because they got the best performance of the lift threshold (see Table 3).

In this work, we considered 3 criteria as our previous work [1]: support, confidence and lift, but this time, we proceed by a process of refinement, which starts by considering only the thresholds of support and confidence to extract set of association rules, then, our process ends with the consideration of the lift measure which allows us to keep just the rules having the best measure of lift.

5.4 Discussion and recommendation

The obtained association rules can be very useful for the decision-maker to do the appropriate strategies according to the conditions most related to road accidents, such as the R6 and R7 rules which show a relation between State of surface and Atmospheric conditions. These association rules indicate and show that the majority of French drivers cause more accidents in normal roads and normal weather conditions. This can be justified by several reasons, the first can be the excess speed in these best atmospheric and road conditions, the second is the increase in driver travel, especially for leisure travel. The third reason is related to the negligence of the driver by the consumption of drugs and the underestimation of the danger of the road.

For the rules R13 and R15 we notice a strong relation between the items: normal surface , on the pavement and normal atmospheric condition, these correlations show that the majority of road accident victims are unfortunately in normal roads with a normal atmospheric. Worse still, most of these pedestrians were hit by vehicles on the pavement. Moreover, hitting a pedestrian on the pavement can only be justified by the imprudence and the carelessness of the drivers. This imprudence returns to several reasons: the very young age of some drivers, the abuse of alcohol and drugs, using the phone while driving, etc.

We note that the results obtained in this present study on road traffic accidents incurred in 2017 are almost the same as those obtained in our work [1] concerning the accidents incurred in 2016.

These association rule needs to be further analyzed with the help of road safety specialist to choose the appropriate policies in the perspective of improving road safety. In conclusion, this knowledge is so relevant for road safety investigators and deserves to be studied further.

6 Conclusion

In this work, we tried to illustrate our contribution about reducing the number of the generated association rules from a large database and considering multiple measures of quality, which allow to the decision maker to select the most interesting rule. To do this we used Apache Spark and machine learning, especially the FP-growth algorithm to extract frequent itemsets and generate association rules. As expected, Apache Spark provided its fast execution engine

For distributed processing. We performed the experiment on road accident data in France in 2017 to extract the main causes of the accidents, by applying FP-growth algorithm in Apache Spark environment to extract the minimum number of association rules with a given minimum support threshold and minimum confidence threshold, and for more facilizing the task of selecting the most useful and relevant rule, we proceeded to a phase of refinement by selecting the rule with a high value of lift measure from the minimal set of rules obtained previously. The obtained rule can be easily exploited by the decision maker to choose the appropriate policies and strategies in the perspective of improving road safety.

For future work, a new methodology should be addressed based on deep learning for a good prediction and finer analysis of road accidents, that may occur in the future.

References

[1] El Mazouri, F.Z., Abounaima, M. & Zenkouar: Data mining combined to the multicriteria decision analysis for the improvement of road safety: case of France. J Big Data (2019) 6:5. https://doi.org/10.1186/s40537-018-0165-0.

[2] Lovell, M. C. (1983): Data Mining. The Review of Economics and Statistics. 65 (1): 112. doi:10.2307/1924403. JSTOR 1924403.

[3] Piatetsky-Shapiro, G., Frawley, W. J.: Knowledge Discovery in Databases. AAAI/MIT Press (1991).

[4] Fayyad, U.M., Piatesky-Shapiro, G., Smuth P., Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press (1996).

[5] Tiwari, A., Gupta, R., Agrawal, D.: A survey on frequent pattern mining: current status and challenging issues. Inf. Technol. J. 9 (2010) 1278–1293.
[6] Agrawal, R., Imielinski, T., Swami A.: Mining association rules between sets of items in large database. In SIGMOD. pp. 207–216 (1993).
[7] Aguinis,H., Forcum, L. E., Joo, J.: Using Market Basket Analysis in Management Research. Journal of management. https://doi.org/10.1177/0149206312466147, (2012).
[8] Zaki, M. J., Sequeira, K.: Data mining in computational biology IN Handbook of Computational Molecular Biology. Chapman & Hall/CRC Press, ch. 38, pp. 1–26 (2006).
[9] Auer, J., Bajorath, J.: Emerging chemical patterns: A new methodology for molecular classification and compound selection, J. Chem. Inf. Mod, vol. 46, no. 6, pp. 2502–2514 (2006).
[10] Neesha, J., Nur'Aini A.R., Wahidah,H.: Data Mining in Healthcare – A Review. Procedia Computer Science. 72 (2015) 306 – 313.
[11] Sajedi,H.: Steganalysis based on steganography pattern Discovery, Journal of information security and applications. 30 (2 0 1 6) 3–14.
[12] http://spark.apache.org. Accessed in 2019.
[13] Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Mining Knowledge Discovery. 2004;8(1):53–87.
[14] Wong, J., Chung, Y.: Comparison of methodology approach to identify causal factors of accident severity. Transp Res Rec 2083:190–198 (2008).
[15] Ait-Mlouk,A, Gharnati, F., Agouti, T.: An improved approach for association rule mining using a multi-criteria decision support system: a case study in road safety. Eur. Transp. Res. Rev. (2017) 9:40. DOI 10.1007/s12544-017-0257-5.
[16] https://www.data.gouv.fr/fr/datasets/base-de-donnees-accidents-corporels-de-la-circulation. Accessed in 2018.
[17] http://www.securite-routiere.gouv.fr/la-securite-routiere/l-observatoire-national-interministeriel-de-la-securite-routiere. Accessed in 2018.
[18] Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proc. 20th Int. Conf. Very Large Data Bases, pp. 487-499 (1994).
[19] Shah, M., Shah, N., Shetty, A., Shah, D., Gotmare, P.: A comparative study of Pattern Recognition Algorithms on Sales Data. International Journal of Computer Applications (0975-8887), Vol 141-No.1 (2016).
[20] Direction de l'information légale et administrative, La sécurité routière en France : Bilan 2017, ONISR, Paris 2018, ISBN : 978-2-11-077443-9.