Ait-Mlouk *et al. J Big Data (2017) 4:42*
DOI 10.1186/s40537-017-0105-4

Journal of Big Data

RESEARCH Open Access

Mining and prioritization of association rules for big data: multi-criteria decision analysis approach

Addi Ait-Mlouk*, Tarik Agouti and Fatima Gharnati

*Correspondence:
aitmlouk@gmail.com
Laboratory of Intelligent Energy Management and Information Systems, Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakech, Morocco

Abstract

Data mining techniques and extracting patterns from large datasets play a vital role in knowledge discovery. Most of the decision makers encounter a large number of decision rules resulted from association rules mining. Moreover, the volume of datasets brings a new challenge to extract patterns such as the cost of computing and inefficiency to achieve the relevant rules. To overcome these challenges, this paper aims to build a learning model based on FP-growth and Apache Spark framework to process and to extract relevant association rules. We also integrate the multi-criteria decision analysis to prioritize the extracted rules by taking into account the decision makers subjective judgment. We believe that this approach would be a useful model to follow, particularly for decision makers who are suffering from conflicts between extracted rules, and difficulties of building only the most interesting rules. Experimental results on road accidents analysis show that the proposed approach can be efficiently achieved more association rules with a higher accuracy rate and improve the response time of the proposed algorithm. The results make clear that the proposed approach performs well and can provide useful information that could help the decision makers to improve road safety.

Keywords: Data mining, Association rules, PFP-growth, Big data, Apache Spark, Road accident, Multi-criteria decision analysis

Introduction

Nowadays, the ultra-connected world is generating massive volumes of data stored in a computer database and cloud environment. These large data need to be analyzed in order to extract useful knowledge and present it to decision makers for further use.

Data mining (non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [1, 2]) techniques are a vital part of many business analytics and predictive applications that come to complete systems that provide prediction techniques and necessary services of analysis. The term of association rules is a powerful technique of data mining for discovering correlation and relationships between objects in the database. It based on statistical analysis and artificial intelligence. This technique is particularly appropriate for analyzing the correlations between objects, because it considers conditional interaction among input data sets, and produce

the decision rules of the form IF-THEN. An example of association rules extracted from a supermarket sales database: "*Cereals* $\wedge$ *Sugar* $\rightarrow$ *Milk* (*Support* 7%, *Confidence* 50%)". This rule means that the customers who buy cereals and sugar also tend to buy milk. The support of the rule is the proportion of customers, who bought the three articles, and the confidence defines as the proportion of customers who bought the milk from those who bought cereals and sugar. Association rules have been used successfully in many areas, including business planning, medical diagnosis, medical research, telecommunications and text mining. Although, the volume of data brings a new challenge to extract patterns such as processing storage and response time for iterative algorithms. Since the datasets are extremely large, parallel algorithms are required. Apache Spark stands as a powerful framework to process and to analyze big datasets by using machine learning algorithms such as FP-growth, this algorithm is a powerful technique for association rules extraction. Apache Spark is considered as the fast and general engine for large-scale data processing due to its speed and scalability. However, its performance decrease in terms of relevance. In addition, it produces a huge number of association rules. Indeed, the final stage of rule validation will let the user face the main difficulty: like the selection of the most interesting rules among the large number of extracted rules. Therefore, it is necessary to help the user in validation task by implementing a preliminary stage of post-processing of extracted rules. The post-processing task aims to reduce the number of rules potentially interesting for the user by using multi-criteria decision analysis. This task must take into account both preferences of decision makers and quality measurement.

The rest of this paper is organized as follows: "Related work" section describes the literature review. "Proposed methodology" section describes the methodology for mining association rules in big data, and its integration with multi-criteria decision analysis to prioritize the extracted rules. The results and discussion are introduced in "Empirical study: road accident analysis" section. In the last section, we concluded by summarizing the work done in the study by providing achievements of the study.

Related work

Data mining techniques are widely used in several research domains and provide useful results to guide the decision makers [3–6]. The results from these techniques, besides being of interest, provide guidance for decision makers. However, in most real cases, it requires many hardware resources to collect and analyze massive data.

Recently, with the rapid expansion of information technologies, data analysis is becoming increasingly complex. To overcome these challenges, many approaches have been proposed. Park et al. [7] proposed an approach based on classification to build a prediction model that can solve the problem of large data in transportation field by using Hadoop MapReduce framework [8]. MapReduce is a great solution for one-pass computations, but not very efficient for use cases that require multi-pass computations. It tends to be slow due to the huge space consumption by each job. Furthermore, Chen et al. [9] proposed an evolutionary algorithm namely Niche-Aided Gene Expression Programming (NGEP) to address the problem of the cost of computing and inefficiency to achieve the goal. Although, these methods have achieved a great deal in obtaining association rules. Moreover, many other research works [10–12] have been proposed but still, suffer from the accuracy and relevance of extracted rules. The motivation of this

research is to propose an approach, which can address this issue, as, mentioned above. In this context, we proposed an approach based on Apache Spark [13] and multi-criteria decision analysis to extract the relevant association rules by using Parallel FP-growth algorithm [14]. This approach is based on four major steps: data preprocessing, frequent patterns extraction without candidate generation, association rules extraction, and the prioritization of extracted rules.

Proposed methodology

In this section, we discuss the various steps that construct our proposed methodology, we started by developing the association rules technique as follows:

Association rules mining

Association rules technique is a powerful data mining method for discovering the relationship between variables in large databases. It was initiated by Agrawal et al. [2], for the first time, to analyze transactional databases. An association rule is usually defined as an implication of the form: $A \rightarrow B$ such as $A, B \subset I$ and $A \cap B = \phi$. Every rule is composed of two different sets of items A and B, where A is called antecedent and B called consequent. For example $\{Driver\} \rightarrow \{Vehicle\}$, suggests that a strong relationship exists between two items $\{Driver,\ Vehicle\}$.

To extract the association rules, two measures are required: the minimum support and the minimum confidence. The support is defined as the proportion of transaction in the database, which contains the items A, the formal definition is:

$$Supp(A \rightarrow B) = Supp(A \cup B) = \frac{|t(A \cup B)|}{t(A)} \tag{1}$$

The confidence determines how frequently items in B appear in transaction that contains A, ranges from 0 to 1, the formal definition is:

$$Confidence(A \rightarrow B) = \frac{Supp(A \cup B)}{Supp(A)} \tag{2}$$

The extraction algorithms of association rules can be classified into three large categories: frequent, maximum and closed algorithms. Many algorithms have been designed to extract frequent itemsets and generate association rules. However, the high number of these algorithms is itself an obstacle to the ability of choice of an expert. In this context, we have done a comparative study [15] based on ELECTRE method to choose the most appropriate algorithms from the large set proposed in the literature. The details of the result are given in Fig. 1.

According to this graph, Apriori-Hybrid and Eclat are incomparable. AprioriTID outranks Partition, Apriori, and Dic. FP-growth outranks Close, and Dic. Dic outrank Relim, Pascal outranks Close and MaxMiner outranks Close. Based on this comparative study, FP-growth is the most appropriate algorithm according to the decision makers' preferences.

FP-growth algorithm

FP-growth algorithm is an efficient and scalable algorithm for mining frequent patterns without using candidate generations, proposed by Han et al. [14]. This algorithm

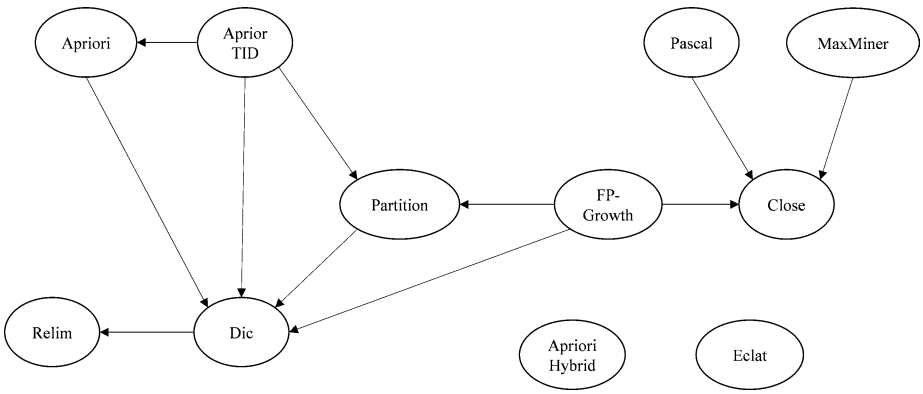

Fig. 1 Outrank graph based on ELECTRE method

encodes the data set using a compact data structure called FP-Tree and extract frequent itemsets directly from this structure. This algorithm has the highest rate of success, it is much faster than Apriori, and use only two passes over datasets. The first step of FP-growth is to compress the input data and mapping each transaction onto the path in the FP tree. When the size of FP-Tree is small, we extract frequent itemsets directly from FP-Tree in memory instead of making multiples pass over the data stored on the disk. Figure 2 shows the dataset with teen transactions, after reading the first transaction the structure of FP-Tree is given in (i) each node contains the label of an item with a counter that shows the number of transactions mapped onto a given path.

Despite its performance, FP-growth algorithm has many drawbacks such as the complexity to extract frequent itemsets for big data and the production of a large number of association rules. To deal with these problems, the adaptation of FP-growth algorithm in big data environment using Apache Spark by integration of multi-criteria decision

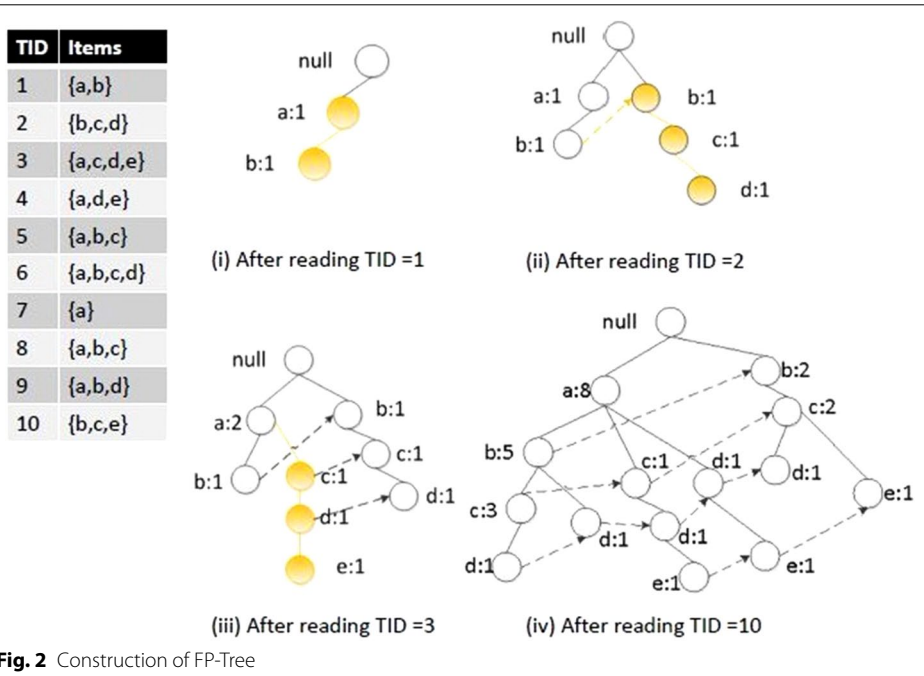

Fig. 2 Construction of FP-Tree

analysis (MCDA) stands as a powerful solution for mining and prioritization of association rules in big data.

Apache Spark

Apache Spark [13] is an open source framework built around speed, ease of use, and sophisticated analytics for big data processing. It was originally developed in 2009 in AMPLab of the University of California, Berkeley, and open sourced in 2010 as an Apache project. Spark takes MapReduce to the next level with less expensive shuffles in the data processing. With capabilities like in-memory data storage and real-time processing, the performance can be several times faster than other big data technologies. Apache Spark runs programs up to 100× faster than Hadoop MapReduce in memory, or 10× faster on the disk. 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.

Spark architecture

Spark applications run on a cluster coordinated by spark context in the main program called driver program. Spark context can connect to several types of cluster managers, once connected spark acquire executors on nodes in the cluster which are processes that run computation and data storage. Afterward, it sends the app to the executors, and finally, spark context sends tasks to the executor to run (Fig. 3).

Multi-criteria decision analysis approach (MCDA)

MCDA is a sub-field of operational research, and management science, dedicated to the development of decision support tools in order to solve complex decision problems involving multi-criteria objectives. Whenever a real decision problem is modeled using multi-criteria analysis, three types of problems are distinguished: choice, sorting and ranking.

In data mining field, the extraction algorithms produce a large number of association rules that not allow the decision makers to make their own choice of the most interesting rules. To deal with this problem the integration of multi-criteria decision

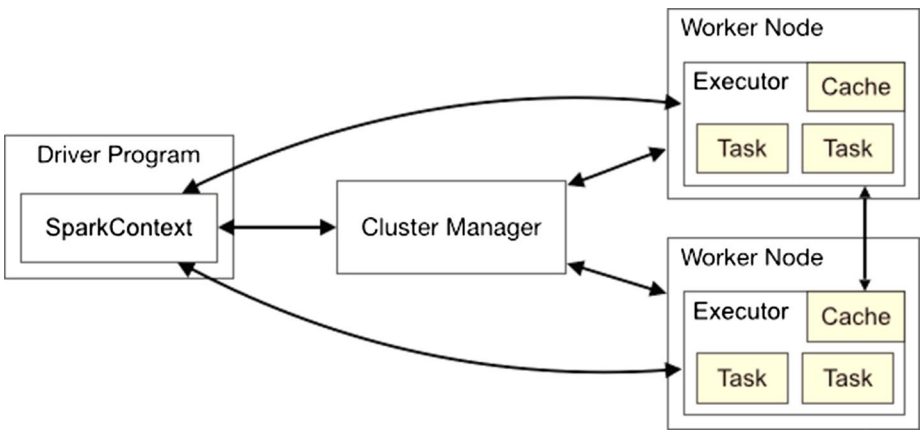

Fig. 3 Spark architecture

analysis, especially an existing method called PROMETHEE, provide the ability to rank the extracted rules according to the proposed relevant measures in the literature [2, 16–21].

PROMETHEE method developed by Brans [22, 23], has been applied in several situations thanks to its ability to simplify and to solve the complex decision problems of ranking type. This method is appropriate to treat the multi-criteria decision problem of the type: $max\{f_1(a),\ldots,f_n(a)|a\,A\}$, by following the given steps:

First of all, it is necessary to determine the matrix of k criteria according to the n different alternatives, Let $A=\{a_1,\ldots,a_n\}$ the set of n alternatives, and $j=\{f_1,\ldots,f_q\}$ the set of q criteria, Table 1.

$$\begin{aligned}\forall a_i,a_j\in A: d_k(a_i,a_j)&=f_k(a_i)-f_k(a_j)\\ \pi(a_i,a_j)&=P_k\left[d_k(a_i,a_j)\right]\end{aligned}\tag{3}$$

The alternative a is better than alternative b according to criteria f, if $f(a)>f(b)$. The preference function can take values on the scale from 0 to 1. When the function of preference has been associated with each criteria by the decision maker, all comparisons between all pairs of actions can be performed for all criteria. A degree of preference is then calculated for each couple of action by the formula (4).

$$\pi(a,b)=\sum_{k=1}^{q}P_k(a,b)\cdot w_k\tag{4}$$

where w_k are weights associated with criteria (close to 1 if very important, close to 0 if very little significant).

The preference matrix is computed by the formula (5).

$$\forall a_i,a_j\in A:\pi(a_i,a_j)=\sum_{k=1}^{q}w_k\pi_k(a_i,a_j)\tag{5}$$

The flow score is computed by the formula (6).

$$\begin{aligned}\phi^{+}(a_i)&=\frac{1}{n-1}\sum_{b\in A}\pi(a_i,b)\\ \phi^{-}(a_i)&=\frac{1}{n-1}\sum_{b\in A}\pi(b,a_i)\\ \phi(a_i)&=\phi^{+}(a_i)-\phi^{-}(a_i)\end{aligned}\tag{6}$$

Table 1 Evaluation table

	F_1	F_2	…	F_k	…	F_q
A_1	$f_1(a_1)$	$f_2(a_1)$	…	$f_k(a_1)$	…	$f_q(a_1)$
A_2	$f_1(a_2)$	$f_2(a_2)$	…	$f_k(a_2)$	…	$f_q(a_2)$
…	…	…	…	…	…	…
A_n	$f_1(a_n)$	$f_2(a_n)$	…	$f_k(a_n)$	…	$f_q(a_n)$

The complete rankings based on the net flow score is given by the formula (7).

$$\begin{aligned} a_i P a_j &\Leftrightarrow \phi(a_i) \succ \phi(a_j) \\ a_i I a_j &\Leftrightarrow \phi(a_i) = \phi(a_j) \end{aligned} \tag{7}$$

The partial ranking based on the positive and negative flow scores is given by the formula (7.1).

$$\begin{aligned} a_i P a_j &\Leftrightarrow \left[\phi^+(a_i) \succ \phi^+(a_j)\right] \wedge \left[\phi^-(a_i) \leq \phi^-(a_j)\right] \\ a_i P a_j &\Leftrightarrow \left[\phi^+(a_i) \geq \phi^+(a_j)\right] \wedge \left[\phi^-(a_i) \prec \phi^-(a_j)\right] \\ a_i I a_j &\Leftrightarrow \left[\phi^+(a_i) = \phi^+(a_j)\right] \wedge \left[\phi^-(a_i) = \phi^-(a_j)\right] \\ a_i J a_j &\quad \textit{otherwise} \end{aligned} \tag{7.1}$$

PROMETHEE GAIA (geometrical analysis for interactive aid) computes the positive and negative preference flows for each alternative, where the positive flow expresses how much an alternative is dominating the other ones, and the negative flow expresses how much an alternative is dominated by the other ones.

GAIA plan (geometrical analysis for interactive aid)
We have:

$$\begin{aligned} \Phi(a_i) &= \frac{1}{n-1} \sum_{b \in A} \sum_{k=1}^{q} w_k \pi(a_i, b) - \frac{1}{n-1} \sum_{b \in A} \sum_{k=1}^{q} w_k \pi_k(b, a_i) \\ &= \sum_{k=1}^{q} w_k \frac{1}{n-1} \sum_{b \in A} \pi_k(a_i, b) - \pi_k(b, a_i) \\ &= \sum_{k=1}^{q} w_k \phi_k(a_i) \end{aligned} \tag{8}$$

Every alternative can be represented by a vector in a space of q dimensions.

$$\vec{\phi}(a) = \left[\phi(a_i), \ldots, \phi_q(a_i)\right] \tag{8.1}$$

Quality measurements

To guide the data analyst identifying interesting rules, many objective measures have been proposed in the literature (Table 2).

Proposed approach

To have an adequate model for discovering association rules from big data, we think it is important to adapt FP-growth algorithm on Apache Spark to PFP-growth (Fig. 4) for mining frequent itemsets then generates association rules. Figure 5 shows the general scheme of the overall algorithm procedure, in this overall system, the association rules mining were generated by the PFP-growth algorithm and the prioritization of extracted rules by PROMEHEE method, the details of the proposed approach are described in Fig. 6 by the following steps:

Preprocessing In this step, we refer to an ETL (extraction transformation loading) tool for preparing and cleansing data by transforming the data to a proper format and selecting only certain columns to load.

Table 2 Quality measurements of association rules

Measure	Formula	
Support The support defined as the proportion of transaction in the database, which contains the items A [2]	$Support(A \rightarrow B) = \frac{\lvert t(A \cup B) \rvert}{t(A)}$	(1)
Confidence The confidence determines how frequently items in B appear in transaction that contains A [2], ranges from 0 to 1	$Confidence(A \rightarrow B) = \frac{Support(A \cup B)}{Support(A)}$	(2)
Lift The lift measures how far from independence are A and B [16]. It ranges within [0, $+\infty$]	$Lift(A \rightarrow B) = \frac{Support(A \cup B)}{Support(A) * Support(B)}$	(9)
Laplace It is a confidence estimator that takes support into account [17]. It ranges within [0, 1]	$lapl(A \rightarrow B) = \frac{Support(A \cup B)+1}{Support(A)+2}$	(10)
Conviction Measure the degree of implication of a rule [18]. It ranges along the values [0.25, $+\infty$]	$conv(A \rightarrow B) = \frac{1-Support(B)}{1-conf(A \rightarrow B)}$	(11)
Leverage Measure how much more counting is obtained from the co-occurrence of the antecedent and consequent from the independence [19]	$leve(A \rightarrow B) = Support(A \rightarrow B) - Support(A) \times Support(B)$	(12)
Jaccard Measure the degree of overlap between the cases covered by each of them [20] the Jaccard coefficient takes values in [0, 1]	$Jacc(A \rightarrow B) = \frac{Support(A \cup B)}{Support(A) + Support(B) - Support(A \cup B)}$	(13)
φ-Coefficient This measure can be used to measure the association between A and B [21]	$\phi(A \rightarrow B) = \frac{leve(A \cup B)}{\sqrt{(Support(A) \times Support(B)) \times (1-Support(A)) \times (1-Support(B))}}$	(14)

Frequent itemsets mining (FIM) Is one of the most intensively investigated problems in terms of computational and algorithmic development. It constitutes the major technique to extract frequent itemsets from datasets.

Association rules mining (ARM) Based on the comparative study [15], we used the FP-growth algorithm to extract association rules adapted (PFP-growth) to the context of big data. The first step of PFP-growth is to compute item frequencies and identify frequent items. The second step uses a suffix tree (FP-Tree) structure to encode transactions without generating candidate sets explicitly. In the final step, the frequent itemsets can be extracted from the FP-Tree by introducing the minimum support (for example, if an item appears 3 out of 5 transactions, it has a support of $3/5 = 0.6$).

Association rules prioritization The process of association rule mining produces a large number of rules that not allow the decision makers to make the right choice of interesting rules. To tackle this problem, we used PROMETHEE method to select only the most interesting rules.

Visualization Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, in order, to that, they can grasp difficult concepts or identify new patterns. In this approach, we used GAIA plan to present the association rules preferences.

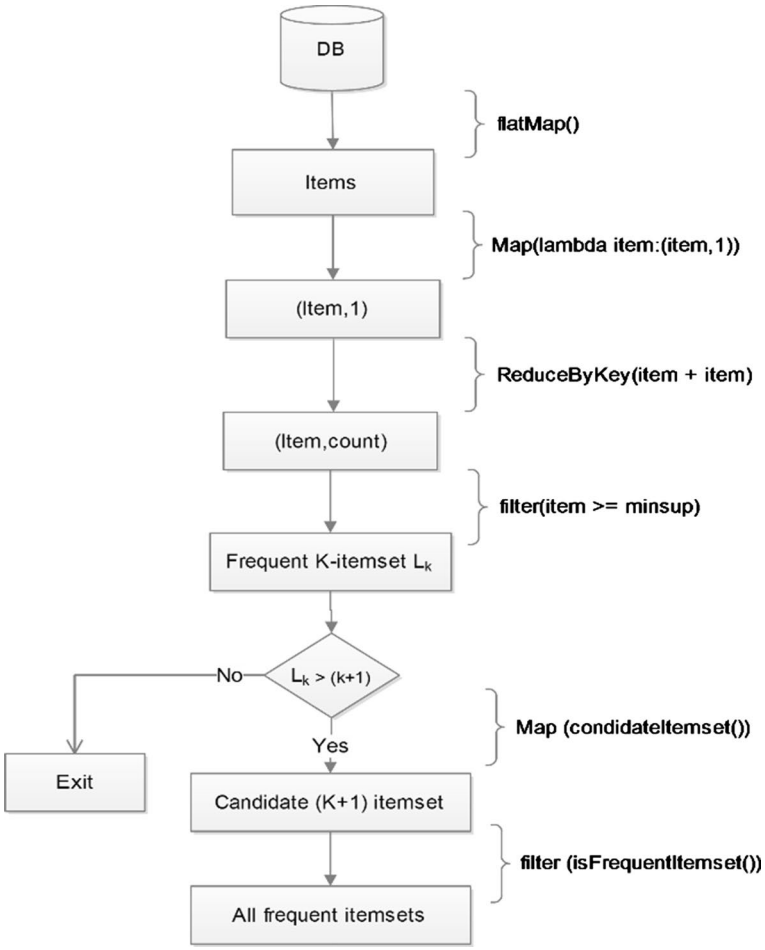

Fig. 4 The overall PFP framework

Empirical study: road accident analysis

Road accidents have emerged as an important public health problem in the world, according to World Health Organization [24], 1.24 million people die in road crashes each year and as many as 50 million are injured. In this study, the accident data were obtained from the Ministry of Equipment, Transport, and Logistics (METL) [25] in the province of Marrakech (Morocco). Moreover, 21 variables were used (Table 3) in order to identify the main factors that affect road accident [26–30]. The variables describe characteristics related to the accident (type and cause), the driver (age, sex, and experience), vehicle (age and type), road (condition and geometry), time, season, number of injuries/death, etc. In addition, the data model used is shown in Fig. 7, it is a CSV file of 200 MB.

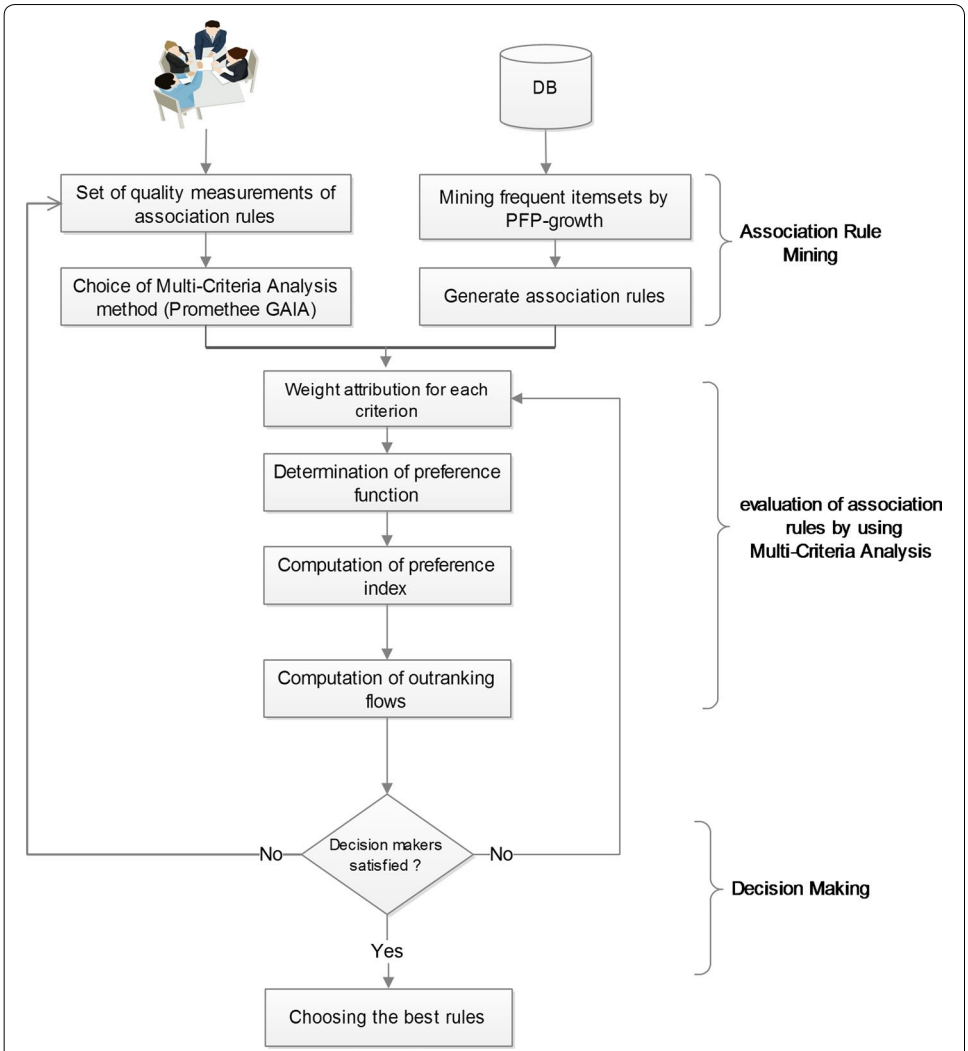

Fig. 5 General algorithm procedure

Results and discussion

At this stage, we present the analysis results of the treatments performed by the association rules, and we end by a multi-criteria analysis to assess the extracted association rules according to the decision makers' preferences. We experiment our proposed approach using road accident data. The system was built on Spark by using single nodes. Our system environment is shown in Table 4.

Frequent itemsets extraction We chose FP-growth the classic algorithm that finds frequent itemsets without candidate generation based on the comparative study [15]. To make FP-growth works on massive data, we used parallel FP-growth version implemented on Apache Spark environment. PFP-growth algorithm takes a Resilient Distributed Dataset (RDD) of transactions, where each transaction is an array of items, Table 5.

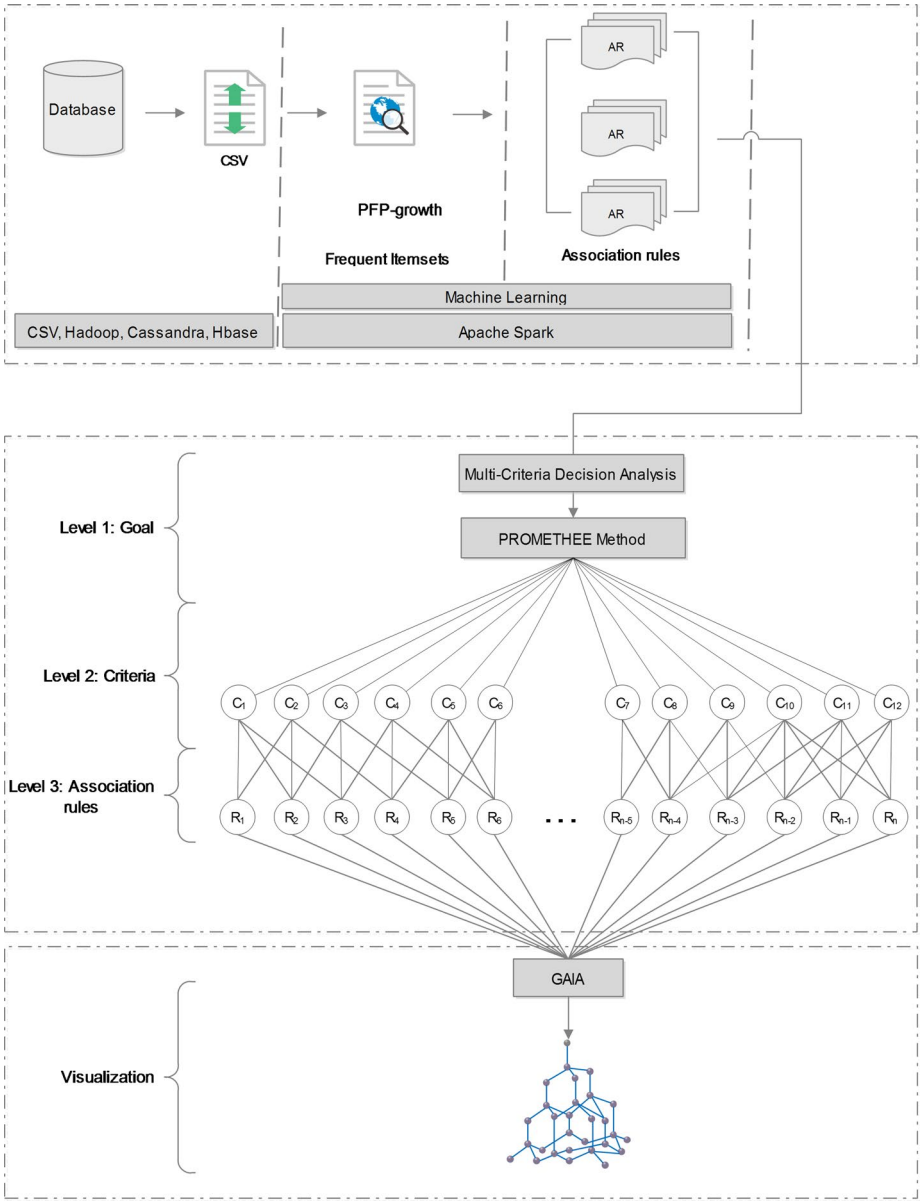

Fig. 6 Steps of proposed approach

Association rules generation This step, implement a parallel algorithm for building association rules that have a single item as the consequent. Two measures are required: the minimum support and the minimum confidence, Table 6.

Data mining algorithms provide a substantial solution for extracting association rules. However, these algorithms produce a large number of rules, which do not allow the decision makers to make their own choice of the most interesting rules. To solve this problem, the integration of multi-criteria decision analysis within the big data platform would be practically useful for the decision makers who are suffering from a large number of

Table 3 Attributes and factors of traffic accident

Attribute name	Values	Description
Accident_ID	Integer	Identification of accident
Accident_Type	Fatal, Injury, Property damage	Accident type
Driver_Age	< 20, [21–27], [28–60], > 61	Driver age
Driver_Sex	M, F	Driver sex
Driver_Experience	< 1, [2–4], > 5	Driver experience
Vehicle_Age	[1–2], [3–4], [5–6], > 7	Service year of the vehicle
Vehicle_Type	Car, Trucks, Motorcycles, Other	Type of the vehicle
Light_Condition	Daylight, Twilight, Public lighting, Night	Light conditions
Weather_Condition	Normal weather, Rain, Fog, Wind, Snow	Weather conditions
Road_Condition	Highway, Ice Road, Collapse Road, Unpaved Road	Road conditions
Road_Geometry	Horizontal, Alignment, Bridge, Tunnel	Road geometry
Road_Age	[1–2], [3–5], [6–10], [11–20], > 20	The age of road
Time	[00–6], [6–12], [12–18], [18–00]	Accident time
City	Marrakesh, Casablanca, Rabat…	Name of city where accident occurred
Particular_Area	School, Market, Shops…	Where the accident occurred
Season	Autumn, Spring, Summer, Winter	Seasons of year
Day	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday	Days of week
Accident_Causes	Alcohol effects, Fatigue, Loss of control, Speed, Pushed by another vehicle, Brake failure	Causes of accident
Number_of_injuries	1, [2–5], [6–10], > 10	Number of injuries
Number_of_deaths	1, [2–5], [6–10], > 10	Number of deaths
Victim_Age	< 1, [1–2], [3–5], > 5	Victim age

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Accident_Type	Drive_Age	Drive_Sex	Drive_Exp	Vehicle_Age	vehicle_Type	Light_Condition	Weather_Condition	Road_Condition	Road_Geometry	Time	Season	Day	Causes
2	Fatal	<20	M	<1	<2	Car	Day	Clear	Collapse road	Horizontal	[6-12]	Spring	Md	Loss of Control
3	Injury	[21-27]	F	>6	<5	Car	Day	Run	Highway	Crossing	[12-18]	Summer	S	Alcohol effects
4	Injury	[28-60]	F	>7	<10	Car	Night	Clear	Collapse road	Alignment	[18-00]	Autumn	W	Speed
5	Injury	>60	F	<1	<15	Car	Day	Run	Highway	Horizontal	[12-18]	Summer	Sa	Speed
6	Injury	<21	F	<2	<10	Truck	Day	Clear	Unpaved road	Alignment	[12-18]	Summer	T	Brake Failure
7	Injury	[21-27]	F	<3	<5	Car	Day	Wind	Highway	Alignment	[6-12]	Winter	Md	Speed
8	Property damage	[28-60]	M	[2-6]	<15	Car	Day	wind	Collapse road	Horizontal	[12-18]	Summer	T	Loss of Control
9	Injury	<21	F	[2-6]	<10	Truck	Day	wind	Unpaved road	Alignment	[12-18]	Autumn	S	Speed
10	Injury	[21-27]	F	[2-6]	<5	Truck	Day	Clear	Highway	Alignment	[12-18]	Summer	W	Pushed by another vehicle
11	Injury	[28-60]	F	[2-6]	<15	Pedestrian	Day	Clear	Collapse road	Crossing	[6-12]	Autumn	Md	Alcohol effects
12	Injury	>61	F	>6	<5	Truck	Day	Clear	Unpaved road	Alignment	[6-12]	Summer	S	Speed

Fig. 7 Data mode

Table 4 Experiment environment

Software	Node environment
Apache Spark 2.0	Single node
Scala IDE 4.4.1	Memory: 12 Gb
SBT 0.13	OS: Ubuntu 16.04 LTS,
	CPU: 2.7 GHz, i7

Table 5 Frequent itemsets

Id	Frequent itemset	Support
1	[Collapse Road]	40
2	[Collapse Road, Clear]	28
3	[Collapse Road, Summer]	27
4	[Collapse Road, M]	35
5	[Collapse Road, M, Day]	28
6	[Collapse Road, Day]	31
7	[[21–27], M]	27
8	[[21–27], Day]	32
9	[Clear]	54
10	[Clear, M]	37
11	[Clear, M, Day]	30
12	[Clear, Day]	45
13	[Horizontal, [21–27]]	27
14	[Horizontal, M]	27
15	[Summer]	52
16	[Summer, Clear]	39
17	[Summer, Clear, M]	27
18	[Summer, Clear, Day]	35
19	[Summer, M]	36
20	[Summer, M, Day]	32
21	[Summer, Car, Day]	30
22	[Summer, Day]	48
23	[S]	38
24	[S, Day]	32
25	[M]	62
…	…	…
27	[M, Day]	47
28	[Car, Clear]	35
29	[Car, Clear, M]	27
30	[Car, Clear, Day]	30
31	[Car, M]	38
32	[Car, M, Day]	29
33	[Car, Day]	41
34	[Unpaved Road]	37
35	[Unpaved Road, Day]	27
36	[Fatal]	42
37	[Fatal, Clear]	33
38	[Fatal, Clear, M]	28
39	[Fatal, Clear, Day]	28
40	[Fatal, Summer]	29

extracted rules. In this approach, the huge number of extracted rules by the PFP-growth algorithm in Spark required the use of multi-criteria analysis ranking method that deals with a large number of alternatives. Therefore, we are interested in an existing method called PROMITHEE by using a set of previously extracted rules as the alternatives to be evaluated according to given criteria in Table 2.

Table 6 The extracted association rules

N	Antecedent	Consequent	Conf
1	Fatal	M	0.85
2	Fatal, Day	Clear	0.90
3	Clear, M	Day	0.81
4	Car, Clear	Day	0.21
5	Fatal, Summer	Clear	0.93
6	[12–18], Summer	Day	0.90
7	Summer, Clear	Day	0.89
8	Clear	Day	0.83
9	3	Car	1.0
10	2, Day	Clear	0.96
11	Collapse Road, Day	M	0.90
12	Summer, Car	Day	0.96
13	Collapse Road, M	Day	0.80
14	[6–12]	Day	0.93
15	[12–18], Day	Summer	0.83
16	< 2, Clear	Day	0.96
17	[21–27]	Day	0.84
18	Summer, M	Day	0.88
19	Summer	Day	0.92
20	S	Day	0.84
21	[12–18]	Day	0.81
22	< 2	Clear	0.90
23	< 2	Car	0.93
24	< 2	Day	0.90
25	Collapse Road	M	0.87
26	Md	M	0.86
27	Fatal, Clear	M	0.84
28	Fatal, Clear	Day	0.84
29	Fatal, Clear	Summer	0.81

Based on the decision makers' preferences, Table 7 gives the evaluation table as a list of values in rows and columns that allow the analyst to identify the performance of relationships between sets of rules and measures. The evaluation table is used to describe a multi-criteria decision analysis problem where each alternative need to be evaluated on N criteria. Moreover, Table 8 represents weights of different criteria used. These weights are non-negative numbers, the higher the weight, the more important the criteria.

The next step is the computation of preference between pairwise (Eq. 4) this function expressing with which degree $Rule_i$ is preferred to $Rule_j$.

Afterward, we compute the partial and global outranking flow (Table 9), then we present the final result of association rules ranking over all criteria used by the decision makers (Table 10). The graphic illustration of the result is obtained by using PROMETHEE GAIA (Fig. 8).

As shown in the results, after the integration of MCDA especially the PROMETHEE method, it is graphically confirmed that rule 12 (Summer, Car $\geq$ Day), has the strongest flow index. Consequently, it is the most relevant one. Eventually, the interesting rules according to the decision makers' preferences are presented in Table 10 from Order 1

Table 7 Evaluation table

Rule\criteria	Support	Lift	Laplace	Confidence	Conviction	Leverage	Jaccard	Phi-coeff
Rule1	36	85	97	2	87	30	50	19
Rule2	31	90	96	2	93	52	55	18
Rule3	37	81	97	2	83	66	65	78
Rule4	35	85	97	2	87	90	36	98
Rule5	29	93	96	3	96	90	45	69
Rule6	33	90	97	2	92	12	89	98
Rule7	39	89	97	2	91	93	45	89
Rule8	54	83	98	1	84	10	65	98
Rule9	29	100	96	3	2	59	43	85
Rule10	30	96	96	3	1	0	45	97
Rule11	31	90	96	2	93	20	16	58
Rule12	31	96	96	3	1	33	99	56
Rule13	35	80	97	2	82	99	89	45
Rule14	32	93	97	2	95	46	78	68
Rule15	36	83	97	2	85	85	98	86
Rule16	30	96	96	3	0	75	69	87
Rule17	38	84	97	2	86	56	68	84
Rule18	36	88	97	2	90	89	94	98
Rule19	52	92	98	1	93	58	59	56
Rule20	38	84	97	2	86	87	93	54
Rule21	44	81	97	1	83	69	98	36
Rule22	30	90	96	3	93	58	97	57
Rule23	30	93	96	3	96	69	94	58
Rule24	30	90	96	3	93	39	98	98
Rule25	40	87	94	2	92	93	93	97
Rule26	36	86	97	2	88	58	89	95
Rule27	33	84	97	2	86	79	97	97
Rule28	33	84	97	2	86	87	36	65
Rule29	33	81	97	2	83	89	91	95

Table 8 Weights of relative importance

Criteria	Support	Lift	Laplace	Confidence	Conviction	Leverage	Jaccard	Phi-coeff
Weight	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

to Order 16. The success of the PROMETHEE method implementation in the process of decision making greatly depends on possibilities and experience of the decision makers.

Due to the huge number of frequent itemsets and extracted rules, we choose only a small set to present and to demonstrate the performance of our proposed approach.

According to the results of the proposed approach, the extracted rules (Table 10) suggest that a strong relationship exists between these attributes {fatal accident, weather conditions, driver sex, road condition, and the season of year}. Fatal accidents tend to occur when the weather is clear, especially in the summer season and when the road condition is collapse road, etc. The theoretical results are identical to those, which are

Table 9 Preference flow

No.	Support	Confidence	Laplace	Lift	Conviction	Leverage	Jaccard	Phi-coeff
Rule1	0.3214	− 0.2500	0.3214	− 0.7857	− 0.2500	− 0.7857	− 0.5714	− 0.8929
Rule2	− 0.4286	0.2857	− 0.6071	0.3929	0.3571	− 0.5714	− 0.5000	− 0.9643
Rule3	0.5000	− 0.8571	0.3214	0.2857	− 0.8571	− 0.1429	− 0.3214	− 0.0357
Rule4	0.1071	− 0.2500	0.3214	0.0000	− 0.2500	0.6071	− 0.8929	0.8214
Rule5	− 0.9643	0.6429	− 0.6071	0.8571	0.6786	0.6071	− 0.7143	− 0.1071
Rule6	− 0.1071	0.2857	0.3214	0.2143	0.1071	− 0.9286	0.0714	0.8214
Rule7	0.7143	0.0714	0.3214	− 0.3214	0.0000	0.7500	− 0.7143	0.3214
Rule8	1.0000	− 0.6786	0.9643	− 1.0000	− 0.7143	− 1.0000	− 0.3214	0.8214
Rule9	− 0.9643	1.0000	− 0.6071	1.0000	1.0000	− 0.2143	1.0000	0.1071
Rule10	− 0.7143	0.8571	− 0.6071	0.8571	0.8571	0.9643	− 0.7143	0.5714
Rule11	− 0.4286	0.2857	− 0.6071	0.3929	0.3571	− 0.8571	− 1.0000	− 0.3571
Rule12	− 0.4286	0.8571	− 0.6071	0.6429	0.8571	0.9643	0.9286	− 0.5714
Rule13	0.1071	− 1.0000	0.3214	− 0.3214	− 1.0000	0.8571	0.0714	− 0.7500
Rule14	− 0.2857	0.6429	0.3214	− 0.1786	0.5714	− 0.6429	− 0.0714	− 0.1786
Rule15	0.3214	− 0.6786	0.3214	− 0.1786	− 0.6429	0.2143	0.7857	0.1786
Rule16	− 0.7143	0.8571	− 0.6071	0.8571	0.8571	0.0714	− 0.1429	0.2500
Rule17	0.6071	− 0.4643	0.3214	− 0.5357	− 0.4643	− 0.5000	− 0.2143	0.0357
Rule18	0.3214	0.0000	0.3214	0.0714	− 0.0714	0.4643	0.4643	0.8214
Rule19	0.9286	0.5000	0.9643	− 0.9286	0.3571	− 0.3571	− 0.4286	− 0.5714
Rule20	0.6071	− 0.4643	0.3214	− 0.5357	− 0.4643	0.3214	0.3214	− 0.6786
Rule21	0.8571	− 0.8571	0.3214	− 0.8571	− 0.8571	− 0.0357	0.7857	− 0.8214
Rule22	− 0.7143	0.2857	− 0.6071	0.5357	0.3571	− 0.3571	0.6071	− 0.4643
Rule23	− 0.7143	0.6429	− 0.6071	0.7143	0.6786	− 0.0357	0.4643	− 0.3571
Rule24	− 0.7143	0.2857	− 0.6071	0.5357	0.3571	− 0.7143	0.7857	0.8214
Rule25	0.7857	− 0.0714	− 1.0000	− 0.7143	0.1071	0.7500	0.3214	0.0000
Rule26	0.3214	− 0.1429	0.3214	− 0.0714	− 0.1429	− 0.3571	0.0714	0.4286
Rule27	− 0.1071	− 0.4643	0.3214	− 0.5357	− 0.4643	0.1429	0.6071	0.5714
Rule28	− 0.1071	− 0.4643	0.3214	− 0.5357	− 0.4643	0.3214	− 0.8929	− 0.2500
Rule29	− 0.1071	− 0.8571	0.3214	0.1429	− 0.8571	0.4643	0.2143	0.4286

provided by the minister of equipment and transport of Morocco, this result may help the decision makers to formulate new policies and strategies for improving road safety.

Previous studies [3, 5, 26, 27, 31–34] have found an association between drivers' behaviors, weather conditions, light conditions, and accident severity. However, the size of database low processing speed which not explored further. The results of our study not only confirm an association between different variables but also show that the integration of PFP-growth algorithm with Apache Spark framework solve the problems of the large dataset and time consumption due to the capabilities of Apache Spark. Moreover, the integration of multi-criteria decision analysis allows the decision makers to extract the relevant association rules.

As explained, the use of the association rules mining algorithms performed well. However, it has several disadvantages, such as time-consuming and more resources are needed, depending on the data size. Consequently, it is very difficult for traditional

Table 10 Interesting rules

Order	Rules	Phi	Phi+	Phi−
1	Rule12	0.3571	0.6384	0.2813
2	Rule18	0.3170	0.6027	0.2857
3	Rule10	0.2768	0.5848	0.3080
4	Rule16	0.2054	0.5580	0.3527
5	Rule7	0.1518	0.5313	0.3795
6	Rule23	0.1250	0.5179	0.3929
7	Rule6	0.1161	0.4911	0.3750
8	Rule24	0.1116	0.4911	0.3795
9	Rule26	0.0714	0.4821	0.4107
10	Rule5	0.0670	0.4911	0.4241
11	Rule4	0.0670	0.4777	0.4107
12	Rule19	0.0670	0.5134	0.4464
13	Rule15	0.0536	0.4732	0.4196
14	Rule25	0.0402	0.4509	0.4107
15	Rule14	0.0402	0.4777	0.4375
16	Rule27	0.0268	0.4464	0.4196
17	Rule29	− 0.0223	0.4330	0.4554
18	Rule22	− 0.0268	0.4286	0.4554
19	Rule20	− 0.0536	0.4107	0.4643
20	Rule9	− 0.1071	0.4196	0.5268
21	Rule8	− 0.1161	0.4241	0.5402
22	Rule17	− 0.1339	0.3750	0.5089
23	Rule3	− 0.1384	0.3839	0.5223
24	Rule21	− 0.1741	0.3616	0.5357
25	Rule13	− 0.2054	0.3527	0.5580
26	Rule2	− 0.2455	0.3304	0.5759
27	Rule28	− 0.2500	0.3080	0.5580
28	Rule11	− 0.2679	0.3170	0.5848
29	Rule1	− 0.3527	0.2768	0.6295

algorithms to deal with big data. Indeed, the integration of association rules analysis technique within Apache Spark solves these problems by improving the processing time (Fig. 9) and the capacity of big data storage (Fig. 10) [35].

The use of PFP-growth over big data produces a large number of association rules. Consequently, it is very difficult to select the most interesting rules. Indeed, the integration of multi-criteria decision analysis approach within the association rules process provides only the significant and interesting rules according to the decision makers' preferences. In conclusion, the proposed approach has the flowing major strengths:

- Manage the complex decision situations by taking into account all the objective and subjective factors.
- Mining interesting association rules for big data.
- Improve the response time for iterative algorithms.
- Improve road safety.

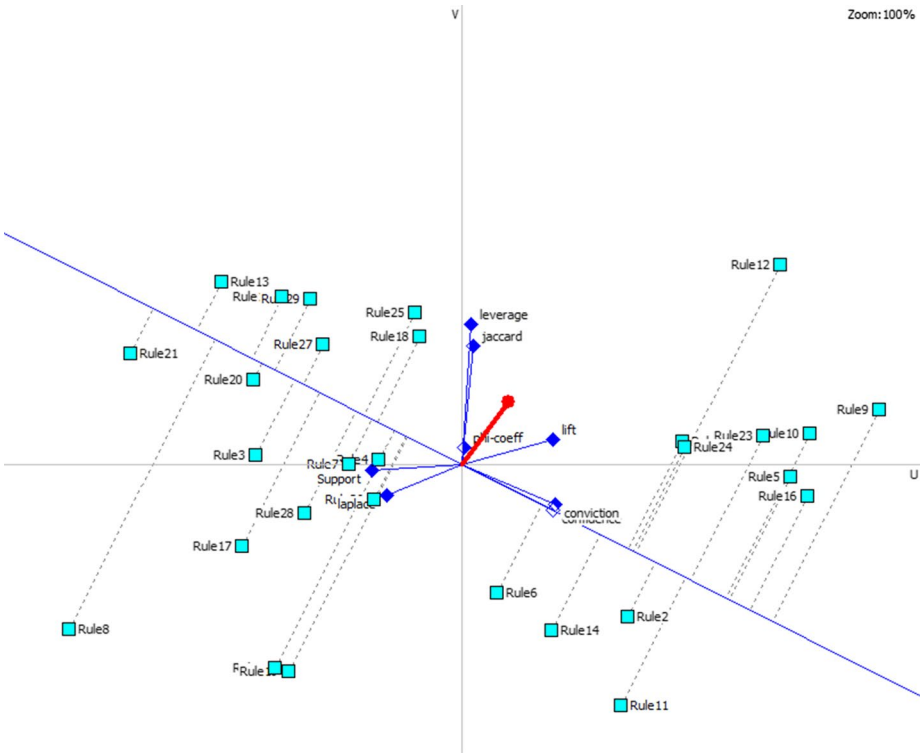

Fig. 8 GAIA plan for association rules ranking

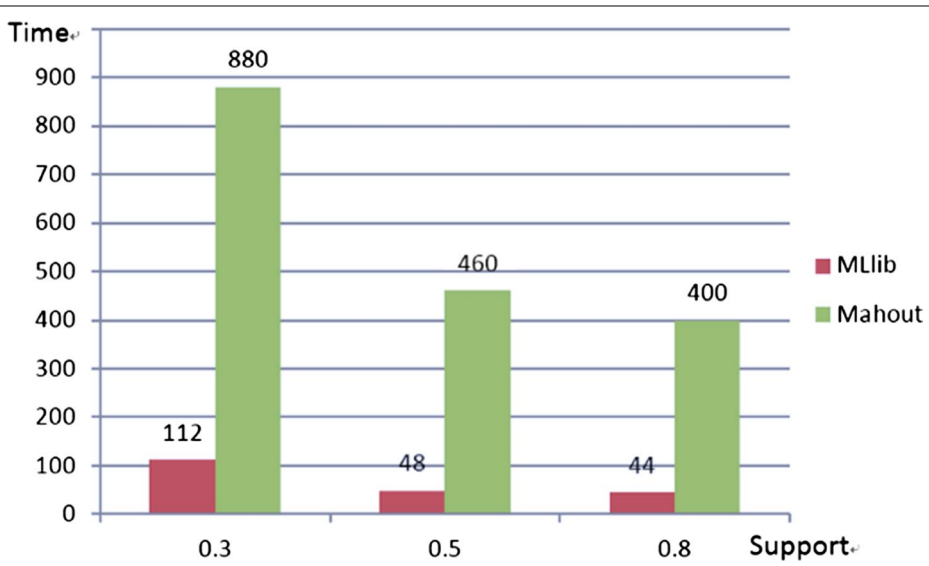

Fig. 9 Running times for different support levels

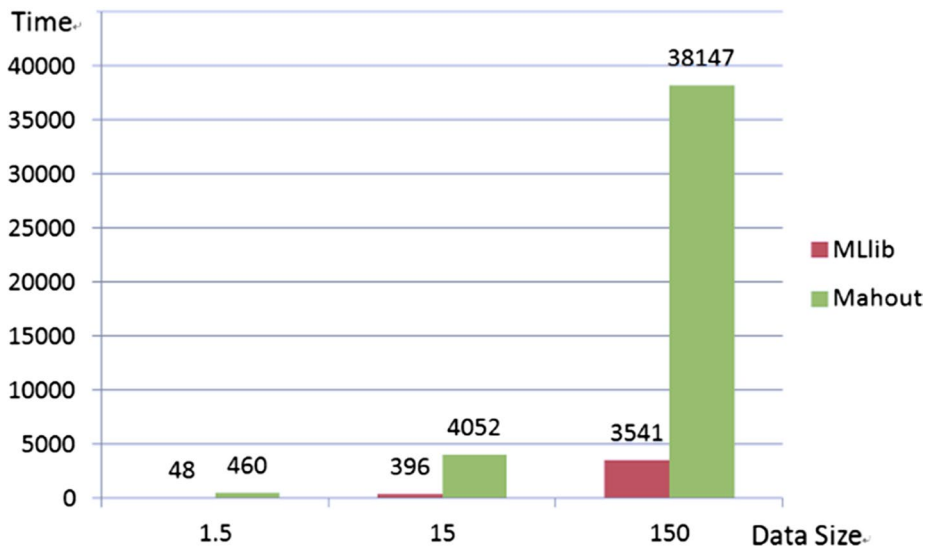

Fig. 10 Running times for different database size

Conclusion

This paper discusses the problem of association rules mining for big data through the road accident. It is clearly identified by using Apache Spark and machine learning, especially PFP-growth algorithm to extract frequent itemsets and generate association rules. Subsequently, we found that Apache Spark provided a faster execution engine for distributed processing and claimed that it is much faster than Hadoop MapReduce as it exploits the advantages of in-memory computations which is particularly more beneficial for iterative computations in the case of iterative algorithms. We performed several experiments on road accident data to measure the speed up and scale up of implementations of our proposed approach. We found out much better than expected results for our experiments. Furthermore, the integration of multi-criteria decision analysis within the association rules mining process solves the problem of a large number of extracted rules by selecting only the most interesting. The results demonstrate that the proposed approach is highly scalable and may assist the decision makers based on some hidden patterns to formulate new rules, strategies, and policies for improving road safety.

For further work, a new methodology should be addressed to process real-time data by using Apache KAFKA the distributed streaming platform, as well as the integration of fuzzy set theory to manage the accuracy of extracted association rules.

Authors' contributions

AAM, as the first author, performed the primary literature review, data collection and experiments, and also drafted the manuscript. TA participated in reviewing and editing the manuscript. FG participated in revising the manuscript. All authors read and approved the final manuscript.

Authors' informations

Addi Ait-Mlouk is a Ph.D. in computer science at the Faculty of Science Semlalia, Cadi Ayyad University, Morocco. He received his Master degree in Computer Science from the Cadi Ayyad University. He is actively engaged in research on various aspects of information technologies ranging from data mining algorithms to big data analytics, fuzzy logic, transportation, multi-criteria analysis, and machine learning.

Fatima Gharnati is an associate professor in the Department of Physics at Cadi Ayyad University, Morocco. Her research interests lie in physics, networks and telecommunication, and embedded systems.

Tarik Agouti is an associate professor in the Department Computer Science at Cadi Ayyad University, Morocco. His research interests lie in mathematical economics, supply chain management, operational research, information systems, decision systems, data mining, SIG, spatial databases, fuzzy logic, multi criteria analysis, and distributed systems.

Acknowledgements
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
Not applicable.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Funding
Not applicable.

Publisher's Note

Received: 23 September 2017 Accepted: 17 November 2017
Published online: 23 November 2017

References

1. Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery: an overview. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, editors. Advances in knowledge discovery and data mining. Cambridge: AAAI Press/MIT Press; 1996. p. 134.
2. Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD conference on management of data (SIGMOD 1993). New York: ACM; 1993. p. 207216.
3. Ossenbruggen P, Pendharkar J, et al. Roadway safety in rural and small-urbanized areas. Accid Anal Prev. 2001;33(4):485–98.
4. Ait-Mlouk A, Agouti T, Gharnati F. An approach based on association rules mining to improve road safety in Morocco. In: International conference on information technology for organizations development (IT4OD). New York: IEEE; 2016. p. 1–6.
5. Sanmiquel L, Rossell JM, Vintro C. Study of Spanish mining accidents using data mining techniques. Saf Sci. 2015;75:49–55.
6. Brenac T. Common before-after accident study on a road site: a low-informative Bayesian method. Eur Transp Res Rev. 2009;1(3):125–34.
7. Park SH, Kim SM, Ha YG. Highway traffic accident prediction using VDS big data analysis. J Supercomput. 2016;72(7):1–17.
8. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. In: Proceedings of the 6th conference on symposium on Operating Systems Design & Implementation-volume 6 (OSDI'04), vol. 6. Berkeley: USENIX Association; 2004. p. 10.
9. Chen Y, Li F, Fan J. Mining association rules in big data with NGEP. Clust Comput. 2015;18:577.
10. Fernandez-Basso C, Dolores M, Martin-Bautista MJ. Extraction of association rules using big data technologies. Int J Des Nat Ecodyn. 2016;11(3):178–85.
11. Prajapati DJ, Garg S, Chauhan NC. Interesting association rule mining with consistent and inconsistent rule detection from big sales data in distributed environment. Future Comput Inform J. 2017;2(1):19–30. https://doi.org/10.1016/j.fcij.2017.04.003.
12. Padillo F, Luna JM, Ventura S. An evolutionary algorithm for mining rare association rules: a big data approach. In: IEEE congress on evolutionary computation (CEC). San Sebastian: IEEE; 2017. p. 2007–2014.
13. http://spark.apache.org/. Accessed 2017.
14. Han J, Pei J, Yin Y, Mao R. Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Min Knowl Disc. 2004;8(1):53–87.
15. Ait-Mlouk A, Agouti T, Gharnati F. Comparative survey of association rule mining algorithms based on multiple-criteria decision analysis approach. In: Control, engineering & information technology (CEIT), 2015 3rd international conference on. New York: IEEE; 2015. p. 1–6, 25–27.
16. Brin S, Motwani R, Silverstein C. Beyond market baskets: generalizing association rules to correlations. In: ACM SIGMOD/PODS 97 joint conference. New York: ACM; 1997. p. 265–276.
17. Good IJ. The estimation of probabilities: an essay on modern Bayesian methods. Cambridge: MIT Press; 1965.
18. Brin S, Motwani R, Ullman JD, Tsur S. Dynamic itemset counting and implication rules for market basket data. In: Peckham J, editor. Proceedings of the 1997 ACM SIGMOD international conference on management of data. New York: ACM; 1997. p. 255–264, 13–15.

19. Piatetsky-Shapiro G. Discovery, analysis, and presentation of strong rules. In: Knowledge discovery in databases. New York: AAAI/MIT Press; 1991. p. 229–248.
20. Jaccard P. Nouvelles recherches sur la distribution florale. Bull Soc Vaud Sci Natl. 1908;44:223–70.
21. Tan P-N, Kumar V, Srivastava J. Selecting the right interestingness measures for association patterns. In: Proceedings of the 2002 ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM; 2002. p. 1–12.
22. Brans JP, Mareschal B, Vincke P. How to select and how to rank projects: the PROMETHEE method. Eur J Oper Res. 1986;24(2):228–38.
23. Brans JP, Mareschal P. The PROMETHEEGAIA decision support system for multi-criteria investigations. Invest Oper. 1994;4(2):107–17.
24. http://www.who.int/gho/road_safety/en/. Accessed 2017.
25. http://www.equipement.gov.ma/en/Pages/home.aspx. Accessed 2017.
26. Kumar S, Toshniwal D. A data mining framework to analyze road accident data. J Big Data. 2015;2:26.
27. Kumar S, Toshniwal D, Parida M. A comparative analysis of heterogeneity in road accident data using data mining techniques. Evol Syst. 2017;8(2):147–55.
28. Kumar S, Toshniwal D. A novel framework to analyze road accident time series data. J Big Data. 2016;3:8.
29. Kumar S, Toshniwal D. Analysis of hourly road accident count using hierarchical clustering and cophenetic correlation coefficient (CPCC). J Big Data. 2016;3:13.
30. Ait-Mlouk A, Gharnati F, Agouti T. Eur Transp Res Rev. 2017;9:40. https://doi.org/10.1007/s12544-017-0257-5.
31. Anderson TK. Kernel density estimation and K-means clustering to profile road accident hotspots. Accid Anal Prev. 2009. https://doi.org/10.1016/j.aap.2008.12.014.
32. Wong J, Chung Y. Comparison of methodology approach to identify causal factors of accident severity. Transp Res Rec. 2008;2083:190–8.
33. Sze NN, Wong SC. Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. Accid Anal Prev. 2007;39:1267–78.
34. Abugessaisa I. Knowledge discovery in road accidents database integration of visual and automatic data mining methods. Int J Public Inf Syst. 2008;1:5985.
35. https://databricks.com/. Accessed 2017.