International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 10 Issue 5 Series I || May 2021 || PP 35-46

Frequent Pattern Analysis of the Roadside Safety Devices Related On-road Crashes

Jianbang Du[1], Fengxiang Qiao[1], Hanzhen Wang[1],Yunpeng Zhang[2], Lei Yu[1]
[1](*Innovative Transportation Research Institute, Texas Southern University, USA*)
[2](*Department of Information and Logistics Technology, University of Houston, USA*)

ABSTRACT:*Roadside safety devices are important in preventing crashes and alleviating crash severities. Their performance criteria that are detailed in the Manual for Assessing Safety Hardware standards, are solely based on full-scale crash testing under ideal site conditions with carefully controlled conditions. The in-service performance measures shall reflect the real functioning of roadside devices, which however has not been well studied. In this paper, the frequent pattern-based data mining approach is adopted to associate the crashes with roadside traffic safety devices. The performances of traffic control devices based on their associated crashes were prioritized to assist in improving the design, test, and maintain roadway traffic control devices for the benefit of safety enhancement. The Apriori and FP-Growth frequent pattern mining algorithms were employed to process ten years' crash data in Texas. Support, confidence, and the evaluation index LIFT were calculated for all cases, while crash severities on "Equivalent Property Damage Only (EPDO)" indexes for each safety device were configured. The frequent pattern mining results imply that the crashes were likely to happen on a dry surface, in clear weather, and under daylight or dark light conditions. The safety device "End of Bridge" was highly associated with the harshest crashes, which suggests that relevant countermeasures and treatments shall be designed and implemented. The device "Side of Bridge" was also related to more severe crashes in earlier years, and is put on a "watch list" for further improvement. "Median barrier" was related to 46% of total crashes, which is however with less EPDO index value.*

KEYWORDS -*Crash Analysis, Data Mining, Frequent Pattern, Traffic Control Devices*

Date of Submission: 07-05-2021 Date of Acceptance: 21-05-2021

I. INTRODUCTION

There were 36,560 nationwide highway fatalities in the year 2018 with a fatality rate of 1.13 per 100 million vehicle miles travels according to the Fatality Analysis Reporting System (FARS) of the U.S. National Highway Traffic Safety Administration (NHTSA) [1]. The major causes of roadway collisions include human factors, vehicle and traffic factors, roadway factors, and environmental factors [2]. Human factors are related to drivers' actions (e.g. speeding) or conditions (e.g., alcohol or drug effects), which contribute the most to crashes, followed by the roadway environment. The roadway factors include roadway design, use of traffic control devices, and land-use configurations [3].

Roadside safety control devices are directly related to roadway design and the use of traffic control devices. They are installed on roadsides to reduce the risk of serious and fatal injuries to motorist's inadvertent road departures. Their performance criteria are detailed in the Manual for Assessing Safety Hardware (MASH) standards, which are however solely based on full-scale crash testing evaluation under ideal site conditions with carefully controlled conditions [4, 5]. The standards have recommended in-service performance evaluation (ISPE) as the final step in evaluating roadside hardware after more than three decades of testing [6]. Differences between field performance and crash test results may appear due to many factors, such as field impact and maintenance conditions, which are not included in crash test guidelines [7]. Impacts of site conditions of safety control devices on crash severity and sensitivity to installation details are not yet well studied. In order to relate crash information with roadside devices, a large amount of data shall normally be processed, which calls for advanced techniques to discover useful knowledge that is embedded in the database.

In this research, the frequent patternbased data mining approach is adopted to characterize the associations between roadside safety devices and different types of crashes. The results of this study will prioritize the performance of traffic control devices based on their associated crashes, so as to improve the design, testing, and maintenance of roadway traffic control devices for the benefits of safety enhancement.

II. LITERATURE REVIEW

Data mining is one of the most practical tools in discovering valuable knowledge from a large number of datasets [8], while frequent pattern mining plays a fundamental role in associating relevancies among

variables. The original concept of frequent pattern mining is the mining of association rules for market-basket analysis[9]. The basic and first frequent pattern algorithm is the Apriori algorithmdeveloped in 1994 based on the generating and testing approach [9]. Another commonly used frequent pattern algorithm is the Frequent Pattern (FP)-growth algorithm proposed in 2004 [10]. The mining of frequent patterns relies on two important measures, which are *Support*and *Confidence* to find frequent itemsets. However, if the thresholds of support are relatively too low or the frequent itemsets are too long, the generated itemsets may not be the most interesting ones [8]. In this case, correlation measures are required to sufficiently filter patterns and generate the most interesting patterns. A typical correlation measure used in frequent pattern mining is called LIFT, which evaluates the correlation between two itemsets by comparing their separate and union occurrences. Such correlation measures can tell whether two itemsets are positively or negatively correlated, so as to improve the performance of frequent pattern mining by narrowing down the data filtering range [11]. Other interesting measures such as χ^2, all confidence, max confidence, Kulczynski, and cosine measures are also applied under different situations.

The frequent pattern algorithms including Apriori and FP-growth have been successfully applied in various scenarios including traffic operation model development. For example, Glatz*et al*. used the frequent pattern mining method in 2014 to visualize traffic network data that contains a large number of communication logs [12]. In a research conducted by Xia *et al.* 2018, a mining method called MapReduce-based Parallel Frequent Pattern growth (MR-PFP) was developed to analyze characteristics in taxi operation [13], which integrated the database, grouped data list, and generated itemsets to find frequent patterns. Another application of frequent pattern analysis is on transportation planning and management of transportation. Juan *et al.* 2008, employed the FP-growth algorithm to process traffic violation data, which was considered as an effective method in the intelligent transportation system [14]. Frequent pattern mining can also be implemented in traffic safety studies. In 2014, Das and Sun [15] employed the association rule method to characterize associations among various factors to discover hidden patterns in rainy weather crash data. Kumar *et al.* 2017, [16] implemented the K-Modes clustering approach to categorize and analyze accident data for heterogeneity reduction. In 2017, Lin *et al.*[17] developed an FP-growth based variable selection method to identify important variables for real-time risk prediction models for traffic accidents.

III. DATA COLLECTION AND METHODOLOGY

3.1 Roadside safety design standards and safety devices

The National Cooperative Highway Research Program (NCHRP) report 350 is known as the early version with evaluating criteria for roadside safety devices [18], while the MASH wasrecently developed with newer criteria. The newest evaluating criteria is MASH 2016 that was implemented in January 2020. It changed some sizes for test vehicles and matricesfor specific roadside safety devices [4]. The list of safety devices in MASH 2016 includes longitudinal barriers, terminals, crash cushions, support structure, work zone attenuation, and channelizers, drainage features, geometric features, and other devices [5]. Roadside safety devices like median barriers are designed and installed on highways to form part of the highway infrastructure, the purpose of which is to reduce the severity of crashes and prevent the occurrence of secondary damages. Typical safety devices installed on highways and roadways include traffic barriers, median barriers, guardrails, bridge rails, and barrier transitions at end of bridges (Fig. 1).

(a) (b) (c)(d) (e) (f)

Figure 1. Typical roadside safety devices on highway and roadway

Fig. 1 demonstrates six types of roadside safety devices. Fig. 1 (a) is a typical subtype of concrete traffic barrier, which is the most common kind of in-service roadside safety device in Texas and some other states. They are installed on divided highways to prevent vehicles from crossing the median and separate opposing traffic. Fig. 1 (b) is an example of low-tension cable median barrier, which was applied for nearly twenty years with similar functions to the concrete media barrier [6]. Fig. 1 (c) is an example of a W-beam guardrail, which is the major type of beam barrier to redirect vehicles that leave the roadway. Bridge rails as shown in Fig. 1 (d) and Fig. 1 (e) are longitudinal barriers, which have the primary function of preventing an errant vehicle from going over the side of the bridge structure that can be categorized into three subtypes: metal railing (Fig. 1 (d)), concrete railing (Fig. 1 (e)), metal and concrete railing [19]. Fig. 1 (f) is a transition of bridge rail end, where the barrier system transits from metal bridge rail to concrete barrier that is often designed as rigid barriers functioning at the side of approaching traffic and adjacent to the traveled way.

3.2 Crash data collection and processing

The crash data used in this study was collected from TxDOT that was maintaining a statewide automated database for received reportable motor vehicle traffic crashes. The crash data were mainly submitted by law enforcement officers with the Texas Peace Officer's Crash Report (form CR-3), submission services, and Crash Records Information System (CRIS). In this study, ten-years (January 1st, 2010 to December 31st, 2019) **5,629,779** crashes were collected. Each of those crash reports is associated with **172** features, including information of crash, unit, person, charges, primary person, endorsements, restrictions, and damages, etc. An extra public specification file was obtained containing description and ID lookup for each type of extracted database. The target variables of this study along with their codes in the crash reports database (Safety Device, Weather Condition, Light Condition, Surface Condition, Day of Week, and Crash Speed Limit) are listed in Table 1.

Table 1. Typical Variables with Codes in Texas Crash Database

Safety Device	Weather Condition	Light Condition	Surface Condition	Day of Week	Crash Speed Limit (mph)
23- guardrail	0- unknown	0- unknown	0- unknown	Monday	-1(No data)
28- work zone barricade, cones, signs or material	2- rain	1- daylight	1- dry	Tuesday	5
	3- sleet/hail	2- dawn	2- wet	Wednesday	10
	4- snow	3- dark, not lighted	3- standing water	Thursday	15
39- median barrier (concrete or cable)	5- fog	4- dark, lighted	5- slush	Friday	20
	6- blowing sand/snow	5- dusk	6- ice	Saturday	25
		6- dark, unknown lighting	8- other	Sunday	30
40- end of bridge (abutment or rail end)	7- severe crosswinds		9- snow		35
	8- other	8- other	10- sand, mud, dirt		…
	11- clear				80
41- side of bridge (bridge rail)	12- cloudy				
56- concrete traffic barrier (not in median)					

The CRIS crash raw data were pre-processed to filter out the desired information and remove the invalid data in order to avoid manipulating the entire huge crash database. During the data preprocessing, the missing or unknown fields were categorized as ID 0.

3.3 Frequent pattern analysis

There are numerous records in the crash data sets, while each is associated with various factors and considered as items in this study. During the data fitting process, the preprocessed data was transformed into a set of lists so that, each crash record including its factors, is an inner list within the outer list of all records. In this way, such records could be directly used as inputs with different items for the processing with frequent pattern mining algorithms [10]. Within the fitted data, each crash record contains a series of binary information of all related items. If a factor is related to a crash record, the corresponding input is one. Otherwise, the input is zero. In this research, the roadside safety devices were considered as one of the items and also records. When the safety devices were considered as part of items, all crash records in the dataset were analyzed as a whole, while all input data were mined together for frequent patterns. When the safety devices were considered separately, the frequent pattern of each safety device was mined accordingly.

According to the theory of data mining, the concept of "pattern" is a set of items, subsequences, or substructures that occur frequently together (or strongly correlated) in a data set. Patterns can represent the intrinsic and important properties of datasets. The process of pattern discovery is to find the inherent regularities in a crash data set. The definitions of several basic concepts are illustrated using a sample set of crash records listed below.

Record 1: Incapacitating injury (I), Median barrier (M), Rain (R), Dusk (D)
Record 2: Incapacitating injury (I), End of bridge (E), Sunny (S), Dusk (D)
Record 3: Fatal injury (F), Median barrier (M), Dusk (U), Tuesday (T)
Record 4: Fatal injury (F), Median barrier (M), Dawn (A), Tuesday (T)
Record 5: Fatal injury (F), Median barrier (M), Sunny (S)

In this example, the term "Item" is the listed attributes on each record (e.g., Fatal Injury, Median barrier, Rain, Dusk…), and the term "Itemset" is a set of one or more items. A k-itemset can be represented as: $X = (x_1, x_2, \dots, x_k)$. The absolute Support or count of X is the frequency or the number of occurrences of itemset X. The relative Support s is the fraction of transactions that contain X, which is also the probability a crash record contains X. An itemset X is frequent if the Support of X is no less than a minsup(minimum support)threshold (σ).

If the minsup = 50%, which means an Item shall appear at least more than (>50%×5 records = 2.5) times, then the frequent 1-itemsets are:

Fatal injury: 3 (3/5=60% > 50%); Median barrier: 3 (3/5=60% > 50); Dusk: 3 (3/5=60% > 50%)

Median barrier: 4 (4/5=80% > 50%)

and, the frequent 2-itemsets are:

{Fatal injury, Median barrier}: 3 (3/5=60% > 50%); {Median barrier, Dusk}: 3 (3/5=60% > 50%)

The association rules would then be: $X \rightarrow Y(S, C)$, where Support s is the probability that a transaction contains $X \cup Y$ (Fig. 2). The Confidence c is the conditional probability that a crash record with both Item X and also Item Y, which is calculated by: $c = sup(X \cup Y)/sup(X)$. The mining of association rule is to find all rules from $X \rightarrow Y$, with the minsup and Confidence. For the above sample case, the association rules with minimum confidence minconf = 50% are:

Fatal injury → Median barrier (60%, 100%)

Median barrier → Fatal injury (60%, 75%)

In these two cases, their Supports are both 60%, but the Confidences c are different (100% vs. 75%). This means, all "Fatal injury" happened on "Median barrier". However, there are other types of injury (actually one "Incapacitating injury" in this example) is also related to "Median barrier".

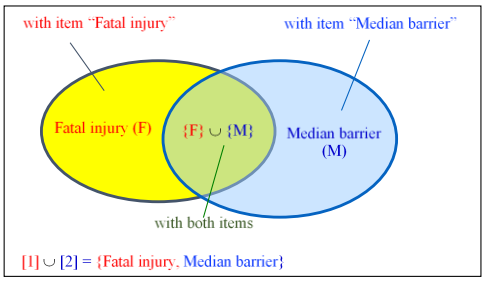


Figure 2. A subtle notation of itemset "$X = Fatal\ injury \cup Y$=Median barrier", which is motivated from [8]

The *Support*, *Confidence*, and the interestingness measurement LIFT can be calculated using (1- 3)[20].

$$s(C, D) = s(C \cup D) = \frac{n(C \cup D)}{n(T)} \quad (1)$$

$$c(C, D) = \frac{s(C \cup D)}{s(C)} \quad (2)$$

$$l(C, D) = \frac{c(C \cup D)}{s(D)} = \frac{s(C \cup D)}{s(C) * s(D)} \quad (3)$$

where,

$s(C, D)$: the *Support* for crash C and device D occurring together, ranging (0, 1);

$n(C, D)$: the number of events when C and D occurring together;

$n(T)$: the number of total events;

$c(C, D)$: the *Confidence* for event D to occur when event C occurs, ranging (0, 1);

$l(C, D)$: the interestingness measurement LIFT (ranging (0, ∞)) for event D to occur when event C occurs, which tells how C and D are correlated;

if $l(C, D) = 1$, events C and D are independent;

if $l(C, D)$ in (1, ∞), events C and D are positively correlated; and

if $l(C, D)$ in (0, 1), events C and D are negatively correlated.

The *Supports*(C, D) can provide the scale of the crash occurring on an influencing item, which is calculated from the number of crashes under the influencing item divide by the total crash number. The *Confidencec*(C, D) is the likelihood of an item occurs if another item happened, which is calculated from the support of two events happen together divide by the support of the single event. The LIFT illustrates the increase in a crash when another item happened, which is calculated from the support of two events that happen together divide by the grade of the supports of the two single events.

The running time, complexity, and quality vary among different algorithms. Thus, the selection of proper frequent pattern algorithms is critical to mining the data sets. Two typical frequent pattern mining algorithms are widely used: the Apriori algorithm and the Frequent Pattern (FP)-Growth algorithm [21]. Fig. 3 is the flow chart and the pseudo-code of the Apriori algorithm.

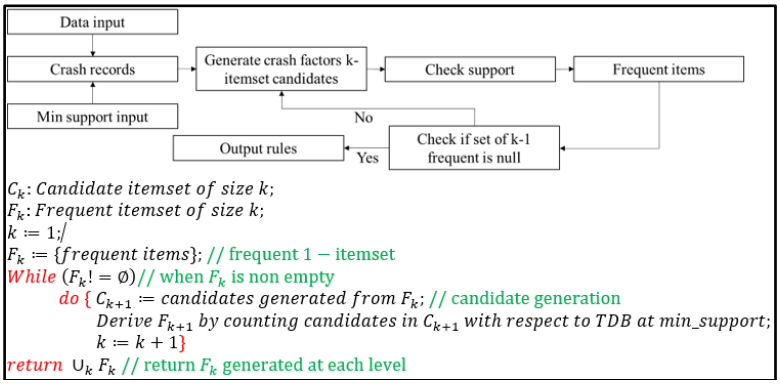


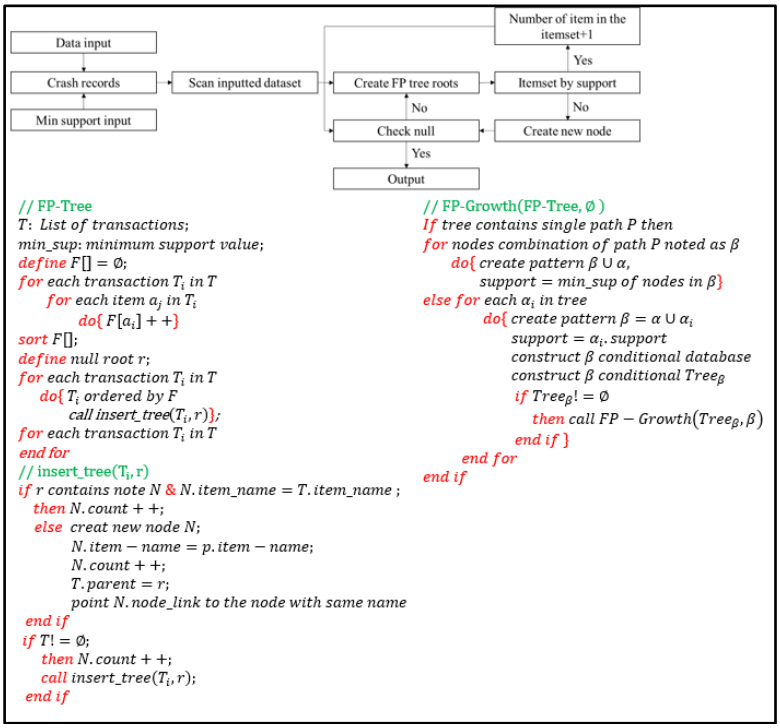
Figure 3. The flow chart and pseudo-code of the Apriori algorithm [8]

The Apriori algorithm scans all possible itemsets and conducts all calculations. As shown in Fig. 3, the itemset candidates of crash factors are generated from the fitted original crash data as inputs, which are compared with the *Support* that is set by the *minsup*. If the *Support* of the candidate itemset is greater than the *minsup*, the frequent items are recorded and the process goes through the null test. The output is then generated after passing null tests. Unlike the Apriori algorithm, the FP-Growth algorithm does not consider all possible itemsets, the flow chart, and pseudo-code of which are illustrated in Fig. 4.

Both theflowchart and the pseudo-code inFig. 4 include two main portions: (1) creating the FP-Tree, and (2) applying the FP-Growth algorithm. As shown in the pseudo-code, the FP-tree is created by: (1) scanning the database once and collecting the dataset *F* along with its *Support* and sort the dataset by descending sequence and saving as a list of datasets; (2) creating the root *r* of the FP-tree and note it as *null*; for each transaction in the database T_i, selecting frequent items and sorting the list, and calling *insert_tree(T_i, r)*; and (3) creating the function *insert_tree(T_i, r)* by checking if node *r* has successive nodes *N* that *N.item-name=p.item-name*, *N* increase by *1* if true, or creating a new node *N* that links to its parent node and set its value to *1*.

To apply the FP-tree and perform the FP-Growth mining, the steps are: (1) checking if *tree* has a single pass *P*, if true, then creating pattern $\beta \cup \alpha$ and setting its *Support* counts as the *minsup* count of β; (2) if false, for each a_i, creating a pattern $\beta = a_i \cup \alpha$ with *support* = a_i.*support*; (3) constructing β conditional tree as $tree_\beta$ and checking if it not *null*, if true, then calling function *FP-Growth($tree_\beta$, β)*. The FP-Growth algorithm only scans the dataset twice when creating the FP-tree for being utilized to store the information [22], which avoids repeated scans in the Apriori algorithm for larger datasets. The inputs of the FP-Growth algorithm include all relevant crash records and the preset *minsup* to be finalized through multiple test runs.

While the mining results of the Apriori and FP-Growth algorithms are the same [23], the FP-growth algorithm runs faster than the Apriori algorithm when the settled *minsup* is under a specific range. If the *minsup* is relatively small, it would be more efficient to use the Apriori algorithm.The original database for the FP-growth can be compressed to decrease the time of the scan [24].



Figurer 4. The flow chart and pseudo-code of the FP-Growth algorithm [25]

To determine the relationship between the roadside safety devices and crash severity, the frequent patterns of the filtered dataset were mined.

The results of this step were the sequences of safety devices by crash severity associated with the *Support*. The reverse frequent pattern mining that obtaining the sequences of crash severity by safety devices was conducted to elaborate and support the results further. To assign a safety index for each roadside safety device, crash severity should be scaled by numeric weights. The Equivalent Property Damage Only (EPDO) weights of crashes are adopted based on the following values [26].

- Scale ID *K*: Fatal injury (death within 30 days), weight 568
- Scale ID *A*: Suspected serious injury, weight 30
- Scale ID *B*: Suspected minor injury, weight 11
- Scale ID *C*: Possible injury, weight 6
- Scale ID *O*: No apparent injury, weight 1

In the EPDO weights, the basis is "1", which demonstrates the average loss of a crash that involves no apparent injury. The crash severity index of each safety device is calculated by (4).

$$Crash\ severity\ index\ = \sum_{r=0}^{r=K}(EPDO_r * s(r)) \quad (4)$$

where: $EPDO_r$: the EPDO weight for severity *r*;
$s(r)$: the support of severity scale ID *r*.

IV. DATA ANALYSIS RESULTS AND DISCUSSION

4.1 Crash trend analysis

Table 2 lists the number of ten-years Texas crashes that were related to different types of roadside safety devices, where the total safety devices related crashes basically increased, except for the drops in 2011 and 2012 (Fig. 5 (left)). Among all safety devices, nearly half (46.0%) crashes were associated with "Median Barriers", while 28.3% and 19.5% crashes were related to "Guardrail" and "Concrete Traffic". Other safety devices were related to the rest of the 6.2% crashes.

Table 2. Number of Ten Years Texas Crash Based on Types of Roadside Safety Devices

Types of Safety Device	Guardrail	Work Zone Barricade, Cones, Signs or Material	Median Barrier	End of Bridge	Side of Bridge	Concrete Traffic	Total Crash
2010	8,586	430	3,098	45	740	8,530	21,429
2011	8,247	426	2,578	20	600	7,996	19,867
2012	8,639	439	4,635	55	625	7,219	**12,973**
2013	7,095	536	13,845	38	1,631	4,110	27,255

2014	6,709	568	15,684	30	1,760	3,490	28,241
2015	7,247	643	18,646	11	1,561	3,781	31,889
2016	7,590	625	17,968	23	986	4,358	31,550
2017	7,903	769	16,246	23	1,049	4,738	30,728
2018	8,171	677	17,759	20	1,106	4,681	32,414
2019	7,786	727	16,114	13	943	4,775	30,358
Total	**77,973**	**5,840**	**126,573**	**278**	**11,001**	**53,678**	**266,704**
	28.3%	**2.1%**	**46.0%**	**0.1%**	**4.0%**	**19.5%**	**100.0%**

The total number of roadside safety devices-related crashes per year are presented in Fig. 5 (right). From the year 2011 to 2014, the number of crashes related to "Median barrier" (the green line) increased from a relatively smaller value to the highest one, while the number of crashes involving "Guardrail" (the blue line) and "Concrete traffic barrier" (the brown line) dropped. Since 2014, the number of crashes related to most types of safety devices slightly fluctuated, and the number of crashes related to "Median barrier" kept the highest, followed by the "Guardrail" and "Concrete barrier".

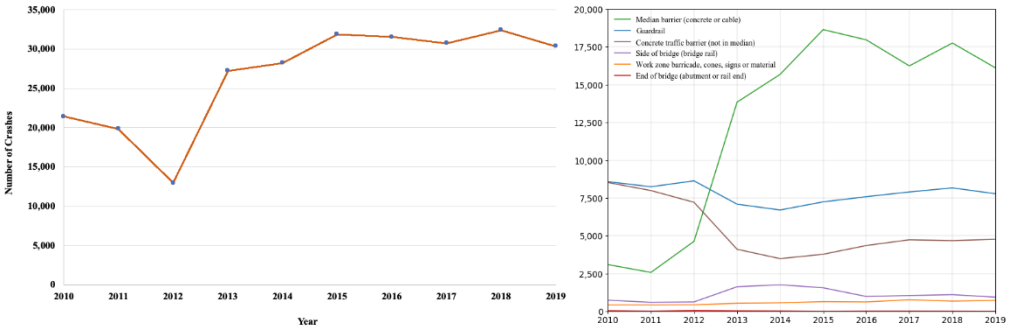


Figure 5. Total number of crashes associated with safety devices in Texas from 2010 to 2019

4.2 Frequent Pattern Analysis

In order to select the suitable algorithm (Apriori or FP-growth) for this research, the average running times under different *minsup* levels were counted in Table 3, which were based on the processing of ten-years' Texas crash data.

Table 3. Running Time of Frequent Pattern Algorithms

minup	**Running time per loop (10 runs, 100 loops each, mean ± std. dev.)**	
	Apriori	**FP-Growth**
60%	127 ms ± 4.61 ms	3.33 s ± 0.155 s
50%	146 ms ± 4.09 ms	4.11 s ± 0.285 s
40%	146 ms ± 2.89 ms	4.48 s ± 0.407 s
30%	156 ms ± 3.67 ms	4.52 s ± 0.709 s
20%	327 ms ± 6.98 ms	6.07 s ± 1.53 s
10%	3.72 s ± 0.111 s	5.84 s ± 0.550 s
5%	14.4 s ± 1.44 s	5.92 s ± 0.312 s
1%	154 s ± 33 s	5.28 s ± 0.614 s

In Table 3, the minsup was set from 1% to 60%. For the cases when minsup > 10%, the Apriori algorithm is superior with less running time. For the cases when minsup is less than 10%, the FP-Growth algorithm is better. Besides, there is a significantly negative relationship between the running time for the Apriori algorithm and the minsup. However, the run time for FP-Growth is relatively stable and not so influenced by the minsup values. Thus, in the rest of this study, the Apriori algorithm was employed when the minsup > 10%. Otherwise, the FP-Growth algorithm was used.

To validate the algorithm accuracy, the Support of each safety device is calculated by the FP-Growth algorithm. Since the proportion of the crashes related to each safety device is greater than 0.01%, the minsup was set as 0.01% in the calculation. The Supports for safety devices as itemsets are:

46.0% for median barrier
28.3% for guardrail
19.5% for concrete traffic barrier
4.0% for side of bridge
2.1% for work zone barricade, cones, signs or material
0.1% for the end of bridge
This result is consistent with the crash trend analysis.

When considering the safety devices as items, the safety devices were considered as items along with other factors such as the "Surface condition", "Day of weeks", "Crash speed limit", "Weather condition", and "Light condition". To identify the minsup, the Support-Confidence plots are shown in Fig. 6.

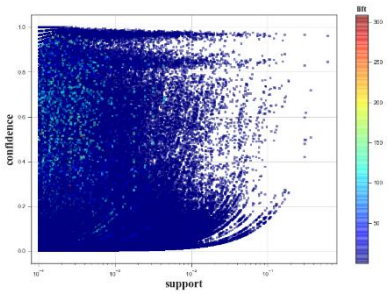


Figure 6. Confidence-Support relationship with the colors representing relevant LIFT values

In Fig. 6, the colored small squares represent the Confidence-Support pairs, while the x-axis is the Support and the y-axis is the Confidence. The right color-bar shows the LIFT, where the warmer color (yellow and red) is related to the bigger LIFT and the colder color (green and blue) means smaller LIFT. The left bottom corner is where the minsup is. In Fig. 6, there is a positive relationship between the Confidences and Supports, which is consistent with the definition of association rules. The mining process generated 568,014 rules with 29,431 frequent itemsets, while 1,915 itemsets have all six items including roadway safety devices and other environmental factors. However, most Supports of these 1,915 itemsets are extremely small due to a large number of datasets. Table 4 shows the itemsets with higher values of Supports.

Table 4. Frequent Itemsets Including Safety Devices

Itemsets						**Support**
Safety Devices	**Weather condition**	**Light condition**	**Surface Condition**	**Day of Week**	**Crash speed limit (mph)**	
Guardrail	Clear	Dark lighted	Dry	Sunday	60	0.15%
Median barrier	Clear	Daylight	Dry	Friday	65	0.25%
Side of bridge	Clear	Daylight	Dry	Wednesday	60	0.01%
	Clear	Daylight	Dry	Tuesday	70	0.01%
Concrete traffic barrier	Clear	Dark lighted	Dry	Sunday	60	0.18%

Table 4 shows a part of the itemsets that contain safety devices and five other environmental factors. The *minsup* was set as 0.01%, while four of the six safety devices are included in the itemsets. The "end of the bridge" (Safety Device ID=40), and the "work zone barricade, cones, signs, or material" (Safety Device ID=28) are excluded in Table 4, which is due to the low occurrence of crashes. The *Support* on each row can be interpreted as the possibility of all events in the occurred itemsets. For example, a hitting guardrail (Safety Devices ID=23) crash is likely to happen under clear weather, under lighted dark condition, on a dry surface, during Sunday, and with a speed limit of 60 mph, with a *Support* of 0.15%. In Table 4, most crashes related to safety devicesare likely to happen in clear weather, on a dry surface, and under daylight. However, since there are safety devicesbeing absent (e.g., "end of bridge" and "work zone barricade"), further frequent mining was necessary to consider each safety device as a "basket" or "record".

When considering the safety devices as separate records, the safety devices are considered separately and the objective of the frequent mining is to find the itemsets and their corresponding *Supports* that appear in each basket (record). Based on the analysis, there are 2,000 to 6,000 rules for each safety device, respectively. The relationships between *Support* and *Confidence* for each safety device are shown in Fig. 7.

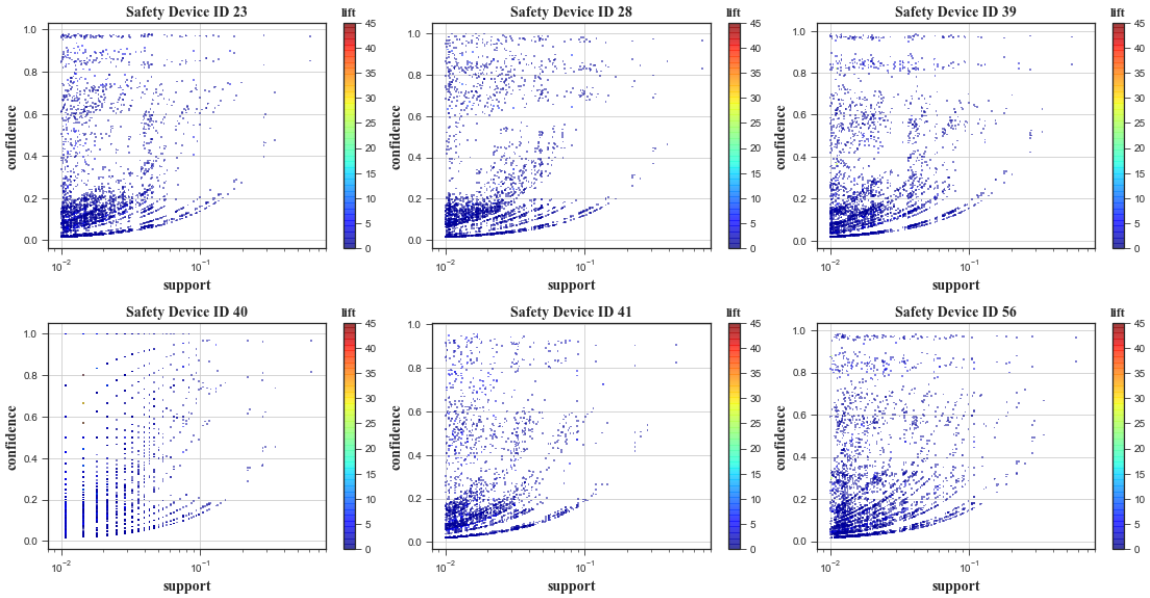


Figure 7. Support-confidence relation of each safety device

In Fig. 7, the colored small squares are the *if-then* association rules of *Confidence* – *Support* pairs for each safety device. The color bar on the right of each subplot represents the value of LIFT, where the warmer colors are with bigger LIFT and the colder colors with smaller LIFT. The *minsup* is at the left bottom corner of each subplot, while the *Confidences* and *Supports* have positive relationships. The higher *Supports* are mainly less than 3% with the *Confidences* less than 20%. Fig. 8 shows the itemsets with higher *Support* for each safety device.

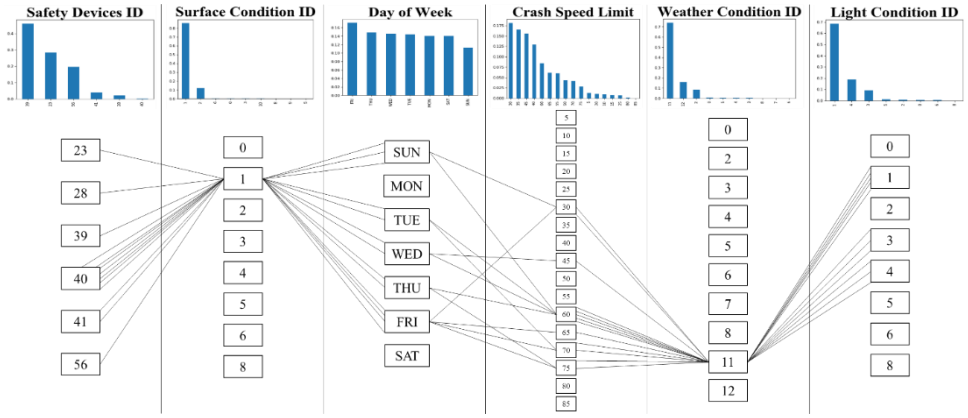


Figure8. Illustration of frequent itemsets with higher *Support* for each Safety Device

The information in Fig. 8 is consistent with Fig. 8 where the safety devices are considered as items. Here, safety devices 28 (work zone barricade, cones, signs or material) and 40 (end of bridge) are now included in the mining, which were however absent in Fig. 6 and Table 4 when considering safety devices as items. The scale of each influencing factor is shown on the top of Fig. 8, where the "dry surface" condition (surface condition ID=1) and "clear weather" condition (weather condition ID=11) have significantly higher appearances than other conditions of these two factors. This could explain the reason why the surface and weather conditions are always the same during frequent pattern analyses. Fig. 8 also illustrates that crashes tended to happen under the speed limits between 30 and 75 mph, with the highest occupancy at 60 mph. Crashes also likely appear more during most of the days of a week, except for Monday and Saturday under daylight (light condition ID =1) or dark (light condition ID =3 and 4) environment.

4.3 Crash severity analysis

The crash severity index with the EPDO weights of each safety device is calculated by Equation 4 using the KABCO system. The prioritized Safety Device ID under different crash severity, the prioritized crash severity under different safety devices, and the EPDO weighted index are displayed in Tables 5, 6, and 7 respectively.

Table 5. The Prioritized Safety Device ID Under Different Crash Severity

Year	Killed/Fatal Injury (K)	Incapacitating Injury/Suspected Serious Injury (A)	Non-Incapacitating Injury (B)	Possible Injury (C)	Unknown/Not Injured (O)
Overall	23, 39, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	23, 39, 56, 28, 41, 40
2019	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 28, 41, 40
2018	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 28, 41, 40
2017	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 28, 41, 40
2016	23, 39, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 28, 41, 40
2015	23, 39, 41, 56, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 28, 41, 40
2014	39, 23, 41, 56, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40
2013	39, 23, 41, 56, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40
2012	23, 39, 56, 41, 28, 40	23, 56, 39, 41, 28, 40	23, 56, 39, 41, 28, 40	56, 23, 39, 41, 28, 40	23, 56, 39, 41, 28, 40
2011	23, 56, 39, 41, 28, 40	23, 56, 39, 41, 28, 40	56, 23, 39, 41, 28, 40	56, 23, 39, 41, 28, 40	23, 56, 39, 28, 41, 40
2010	23, 56, 39, 41, 40, 28	23, 56, 39, 41, 28, 40	56, 23, 39, 41, 28, 40	56, 23, 39, 41, 28, 40	23, 56, 39, 41, 28, 40

Table 5 prioritizes roadside safety devices under different levels of crash severity. In the table, it is shown that for "killed/fatal injury K" and for "unknown / not injury O", the guardrail (Safety Device ID=23) ranks number one of all safety devices, while the end of bridge (Safety Device ID=39) ranks the last. For all other levels of crash severities, the safety device median barrier (Safety Device ID=23) ranks number one. For the separated analysis among the ten years, the ranks after and before 2013 are different, which is consistent with the trends of data counting.

Table 6. The Prioritized Crash Severity Under Different Safety Device

Year	Guardrail (23)	Work Zone Barricade, Cones, Signs or Material (28)	Median Barrier (39)	End of Bridge (40)	Side of Bridge (41)	Concrete Traffic Barrier (56)
Overall	O, C, B, A, K	O, C, B, A, K	O, C, B, A, K	O, **B, C, K, A**	O, C, B, A, K	O, C, B, A, K
2019	O, C, B, A, K	O, C, B, A, K	O, C, B, A, K	O, **K, C, B, A**	O, C, B, A,K	O, C, B, A, K
2018	O, C, B, A, K	O, C, B, A, K	O, C, B,A, K	O, **C, B, K, A**	O, C, B, A, K	O, C, B, A, K
2017	O, C, B, A, K	O, C, B, A, K	O, C, B, A, K	O, **C, B, K, A**	O, C, B, A,K	O, C, B, A, K
2016	O, C, B, A, K	O, C, B, A, K	O, C, B,A, K	O, **K, B, A, C**	O, C, B, A, K	O, C, B, A, K
2015	O, C, B, A, K	O, C, B, A, K	O, C, B,A, K	O, **C, K, B, A**	O, C, B, A, K	O, C, B, A, K
2014	O, C, B, A, K	O, C, B, A, K	O, C, B,A, K	O, **B, K, C, A**	O, C, B, A,K	O, C, B, A, K
2013	O, C, B, A, K	O, C, B, A, K	O, C, B, A, K	O, **B, C,** A, K	O, C, B, A, K	O, C, B, A, K
2012	O, C, B, A, K	O, C, B, A, K	O, C, B,A, K	O, **K, C,B, A**	O, **B, C,** A, K	O, C, B, A, K
2011	O, C, B, A, K	O, C, B, A, K	O, C, B,A, K	O, **B, A, K, C**	O, C, B, **K, A**	O, C, B, A, K
2010	O, C, B, A, K	O, C, B, A, K	O, C, B,A, K	O, **B, K, C, A**	O, C, B, **K, A**	O, C, B, A, K

The prioritized levels of crash severities for each roadside safety device in Table 6 illustrates that, for most safety devices in most years, the prioritized levels of crash severities are: **O, C, B, A, K**, an exactly inverse order of crash severity. This makes sense since normally severer crashes shall be less happened than no severer crashes for particular roadside safety devices. However, the sequences of crash severities for the safety device "end of bridge" in all years are quite different from those for other safety devices. Especially, the crash severity level "killed/fatal injury (K)" for "end of bridge" ranked number 4 (10 years overall; and years 2018, 2017, and 2011), number 3 (years 2015, 2014, and 2010), and even number 2 (for years 2019, 2016, and 2012). This implies that certain countermeasures and treatments shall be designed and implemented nearby the device "end of bridge" in Texas so as to reduce the severity level of crashes.Another special roadside device is the "side of bridge", the crash severity sequences of which in years 2010-2012 were different from most of the other years and safety devices. Especially in the years 2010 and 2011, the crash severity level "killed/fatal injury (K)" ranked number four instead of number five. As such situations totally improved since the year 2013, the safety device "side of bridge" can be put on a "watch list" with no immediate actions of treatments.

Table 7. The Safety Device / Crash Severity EPDO Index

Year	Guardrail	Work Zone Barricade, Cones, Signs or Material	Median Barrier	End of Bridge	Side of Bridge	Concrete Traffic Barrier
Overall	11.06	10.74	7.97	**68.31**	18.18	7.72
2019	9.11	8.74	7.74	**136.77**	15.81	6.97
2018	11.25	7.03	7.74	33.15	14.10	8.17
2017	10.81	8.09	7.65	29.04	15.44	8.43
2016	**12.52**	**16.91**	7.41	129.17	18.65	**8.70**
2015	12.28	7.01	7.00	54.82	17.17	7.99
2014	11.59	12.40	7.55	43.70	13.05	7.97
2013	10.68	13.06	8.97	51.66	19.29	7.91
2012	10.13	15.87	**11.76**	85.20	25.47	6.45
2011	10.67	9.90	11.61	5.35	26.93	7.65
2010	11.76	11.79	10.14	94.20	**29.14**	7.80

Table 7 provides quantitative measures of the relationship between roadside safety devices and on-road crashes with the EPDO weighted index. The interpretation of the EPDO index is the number of equivalent no-injury crashes. Higher values of such index mean even severer. For example, in the year 2016, the loss of a crash relative to "guardrail" equals the loss of 12.52 no-injury crashes. In the year 2019, the loss of a crash relative to the "work zone barricade, cones, signs or material" equals the loss of 8.74 no-injury crashes. Again, the roadside safety device "end of bridge" is with the highest EPDO index (68.31 marked in yellow color) of all devices for all ten years of data. The gray-colored bold cells show the highest crash severity values of the relevant safety devices of each year.

In Table 7, the device "end of bridge" is listed as the highest EPDO weighted indexes of all roadside safety devices, especially for the years 2019, 2016, and 2010. This again suggests the attention on "end of bridge" for immediate actions of safety enhancement. The "concrete traffic barrier" and the "median barrier" are with the lowest EPDO index, and there is no significant trend suggested over the ten years. From the previous analysis, the "median barrier" is the most crash-related roadside safety device, which counts for a large share (46%) of the total number of crashes. However, with the crash severity being considered, the "median barrier" is less significant than the "end of bridge" and the "side of bridge", which is because that some "median barriers" are located within relatively wider medians with more buffer areas than other safety devices. In the meantime, the number of crashes related to the "end of bridge" and the "side of bridge" is less but normally severer, which is consistent with the suggestions from Tables 5 and 6.

V. CONCLUSION

This paper applies the Apriori and FP-Growth frequent pattern mining algorithms to characterize the relationship between roadside safety devices and on-road crashes. The flow chart and pseudo-codes of the Apriori and FP-Growth algorithms and the calculation equations of the evaluation parameters were provided. Ten-year roadway crash data from TxDOT database were collected, which contains various crash influencing factors and six target roadside safety devices. The raw data were fitted into a set of lists as part of the input, along with the minimum supportfor the frequent pattern algorithm. The proper algorithm was selected based on the optimal running time. Through data analysis, the trends of the ten-year crash data were depicted. The associations between the crashes and their influencing factors were elaborated through the mining of frequent patterns.

The frequent pattern mining results suggest that crashes are likely to happen on dry surface pavement, in clear weather, and under daylight or dark light conditions. No-injury crashes rank number one for all roadside devices, while fatal crashes rank the last for roadside safety devices except for the "end of bridge" and the "side of bridge". It is suggested that certain countermeasures and treatments shall be designed and implemented for the roadside device "end of bridge", while the "side of bridge" shall be put on a "watch list". Besides, the EPDO weights are adopted for a crash severity index. The crash severities did not vary much within the 10 years. The mildest crashes were related to the "concrete traffic barrier", and the harshest crashes were related to the "end of bridge". The average crash severities related to the roadside safety devices were likely to be severer than a suspected minor injury. The safety device "media barrier", while is related to 46.0% of the total crashes in Texas, the EPDO index of which is however generally very low.As a plan of the future work of this study, the design (color, reflection, etc.), length, year of service, and maintenance records of roadside devices will be included in the next phase studies.

ACKNOWLEDGEMENTS

The authors acknowledge that this research is supported in part by the Texas Department of Transportation (TxDOT) with project number 0-7018. The authors appreciate the comments and support from the project manager Wade Odell and other members of the Project Management Committee. The opinions,

findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] US DOT, T.S.F. *NHTSA's National Center for Statistics and Analysis*. 2019.

[2] Kiran, B.N., N. Kumaraswamy, and C. Sashidhar, *A review of road crash prediction models for developed countries.* American journal of traffic and transportation engineering, 2017. **2**(2): p. 10-25.

[3] Ossenbruggen, P.J., J. Pendharkar, and J. Ivan, *Roadway safety in rural and small urbanized areas.* Accident Analysis & Prevention, 2001. **33**(4): p. 485-498.

[4] AASHTO, *Manual for assessing safety hardware (MASH)*. 2009, Washington, DC.

[5] AASHTO, *Manual for assessing safety hardware (MASH)*. 2016, Washington, DC.

[6] Cooner, S.A., et al., *Performance evaluation of cable median barrier systems in Texas*. 2009.

[7] Schalkwyk, I.V., et al. *An Overview of the Development of a TxDOT In-Service Performance Evaluation Process for Roadside Safety Features*. 2006.

[8] Han, J., J. Pei, and M. Kamber, *Data mining: concepts and techniques*. 2011: Elsevier.

[9] Agrawal, R., T. Imieliński, and A. Swami. *Mining association rules between sets of items in large databases*. in *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*. 1993.

[10] Han, J., et al., *Mining frequent patterns without candidate generation: A frequent-pattern tree approach.* Data mining and knowledge discovery, 2004. **8**(1): p. 53-87.

[11] Hussein, N., A. Alashqur, and B. Sowan, *Using the interestingness measure lift to generate association rules.* Journal of Advanced Computer Science & Technology, 2015. **4**(1): p. 156.

[12] Glatz, E., et al., *Visualizing big network traffic data using frequent pattern mining and hypergraphs.* Computing, 2014. **96**(1): p. 27-38.

[13] Xia, D., et al., *A MapReduce-based parallel frequent pattern growth algorithm for spatiotemporal association analysis of mobile trajectory big data.* Complexity, 2018.

[14] Juan, X., et al. *Association rule mining and application in intelligent transportation system*. in *2008 27th Chinese Control Conference*. 2008. IEEE.

[15] Das, S. and X. Sun. *Investigating the pattern of traffic crashes under rainy weather by association rules in data mining*. in *Transportation Research Board 93rd Annual Meeting*. 2014. Washington, DC.

[16] Kumar, S., D. Toshniwal, and M. Parida, *A comparative analysis of heterogeneity in road accident data using data mining techniques.* Evolving systems, 2017. **8**(2): p. 147-155.

[17] Lin, L., Q. Wang, and A.W. Sadek, *Real-time traffic accident risk prediction based on frequent pattern tree.* arXiv preprint arXiv, 2017.

[18] Ross Jr, H.E., et al., *Recommended Procedures for the Safety Performance Evaluation of Highway Features*. 1993: Washington, DC.

[19] TxDOT, *Bridge Railing Identification Guide*. May, 2020.

[20] Lin, L., Q. Wang, and A.W. Sadek, *A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction.* Transportation Research Part C: Emerging Technologies, 2015. **55**: p. 444-459.

[21] Aggarwal, C.C., M.A. Bhuiyan, and M. Al Hasan, *Frequent pattern mining algorithms: A survey*, in *Frequent pattern mining*. 2014, Springer: Cham. p. 19-64.

[22] Koh, J.L. and S.F. Shieh. *An efficient approach for maintaining association rules based on adjusting FP-tree structures*. in *International Conference on Database Systems for Advanced Applications*. 2004. Springer, Berlin, Heidelberg.

[23] Xin, D., et al. *Mining compressed frequent-pattern sets*. in *Proceedings of the 31st international conference on Very large data bases*. 2005.

[24] Kavitha, M. and S.T. Selvi, *Comparative Study on Apriori Algorithm and Fp Growth Algorithm with Pros and Cons.* International Journal of Computer Science Trends and Technology (I JCS T), 2016. **4**.

[25] Li, H., et al. *Pfp: parallel fp-growth for query recommendation*. in *Proceedings of the 2008 ACM conference on Recommender systems*. 2008.

[26] Harmon, T., G.B. Bahar, and F.B. Gross, *Crash Costs for Highway Safety Analysis*. 2018.

Jianbang Du, et. al, "Frequent Pattern Analysis of the Roadside Safety Devices Related On-road Crashes." *International Journal of Engineering Science Invention (IJESI)*, Vol. 10(05), 2021, PP 35-46. Journal DOI-10.35629/6734