Profiling high frequency accident locations using associations rules

RA-2002-02

Karolien Geurts, Geert Wets, Tom Brijs, Koen Vanhoof

Onderzoekslijn kennis verkeersonveiligheid

DIEPENBEEK, 2012.
STEUNPUNT VERKEERSVEILIGHEID BIJ STIJGENDE MOBILITEIT.

Documentbeschrijving

Rapportnummer:	RA-2002-02
Titel:	Profiling high frequency accident locations using association rules
Auteur(s):	Karolien Geurts, Geert Wets, Tom Brijs, Koen Vanhoof
Promotor:	Geert Wets
Onderzoekslijn:	kennis verkeersonveiligheid
Partner:	Limburgs Universitair Centrum
Aantal pagina's:	18
Trefwoorden:	high frequency accident locations, black spot, data mining, association rules, describing accidents.
Projectnummer Steunpunt:	1.2
Projectinhoud:	Analyse en detectie van zwarte zones

Uitgave: Steunpunt Verkeersveiligheid bij Stijgende Mobiliteit,december 2002.

Steunpunt Verkeersveiligheid bij Stijgende Mobiliteit
Universitaire Campus
Gebouw D
B 3590 Diepenbeek

T 011 26 81 90
F 011 26 87 11
E info@steunpuntverkeersveiligheid.be
I www.steunpuntverkeersveiligheid.be

SAMENVATTING

In België behoort verkeersveiligheid momenteel tot een van de belangrijkste prioriteiten van de regering. Het identificeren en profileren van zwarte punten en zwarte zones in termen van ongevallendata en locatie kenmerken moet dan ook nieuwe inzichten bieden in de complexiteit en oorzaken van verkeersongevallen. Deze inzichten moeten vervolgens een waardevolle input zijn voor beleidsacties ten behoeve van de verkeersveiligheid. In deze paper worden associatieregels gebruikt om ongevalomstandigheden te identificeren die vaak samen voorkomen op locaties met een hoge ongevallenfrequentie. Vervolgens worden deze patronen geanalyseerd en vergeleken met vaak voorkomende patronen op locaties met een lage ongevallenfrequentie. De sterkte van deze benadering ligt in de identificatie van de relevante variabelen die een belangrijke bijdrage leveren in het begrijpen van ongevallen en hun omstandigheden en in het onderscheiden van beschrijvende ongevalpatronen van de ongevalpatronen die discriminerend zijn voor hoge en lage ongevallenfrequentie locaties om op die manier zwarte punten en zwarte zones te profileren. Het gebruik van dit data mining algoritme is vooral nuttig in de context van grote datasets van verkeersongevallen gezien dat data mining omschreven kan worden als de extractie van informatie uit grote hoeveelheden data. Resultaten tonen aan dat het menselijk aspect en gedrag van groot belang zijn wanneer we frequent voorkomende ongevalpatronen analyseren. Deze factoren spelen ook een belangrijke rol in het identificeren van verkeersveiligheidsproblemen in het algemeen. Maar de meest discriminerende ongevalkarakteristieken tussen hoge en lage ongevallenfrequentie locaties zijn vooral gerelateerd aan infrastructuur en locatie kenmerken.

ABSTRACT

In Belgium, traffic safety is currently one of the government's highest priorities. Identifying and profiling black spots and black zones in terms of accident related data and location characteristics must provide new insights into the complexity and causes of road accidents which, in turn, provide valuable input for government actions. In this paper, association rules are used to identify accident circumstances that frequently occur together at high frequency accident locations. Furthermore, these patterns are analysed and compared with frequently occurring accident characteristics at low frequency accident locations. The strength of this approach lies within the identification of relevant variables that make a strong contribution towards a better understanding of accident circumstances and the discerning of descriptive accident patterns from more discriminating accident circumstances to profile black spots and black zones. The use of this data mining algorithm is particularly useful in the context of large datasets on road accidents, since data mining can be described as the extraction of information from large amounts of data. Results show that human and behavioural aspects are of great importance when analysing frequently occurring accident patterns. These factors play an important role in identifying traffic safety problems in general. However, the most discriminating accident characteristics between high frequency accident locations and low frequency accident locations are mainly related to infrastructure and location characteristics.

INTRODUCTION

In Belgium, every year approximately 50.000 injury accidents occur in traffic, with almost 70.000 victims, of which 1.500 deaths. In 1998 the probability of having a deadly accident (relatively to the number of vehicle-kilometers travelled) was almost 35% higher in Belgium than the European average. Based on these figures, Belgium has a bad record towards traffic safety in comparison with most other European countries (*1*). Not only does the steady increase in traffic intensity pose a heavy burden on the society in terms of the number of casualties, the insecurity on the roads will also have an important effect on the economic costs associated with traffic accidents. Accordingly, traffic safety is currently one of the highest priorities of the Belgian government.

Since a few decades, traffic accident data are registered and analysed to support the traffic safety policy. The identification of geographical locations with highly concentrated traffic accidents (black spots and black zones) and profiling them in terms of accident related data and location characteristics must therefore provide new insights into the complexity and criteria that play a significant role in the occurrence of traffic accidents to provide valuable input for government actions towards traffic safety. According to Kononov (*2*) it is not possible to develop effective counter-measures without being able to properly and systematically relate accident frequency and severity to a large number of variables such as roadway geometries, traffic control devices, roadside features, roadway conditions, driver behaviour or vehicle type.

Lee, Saccomanno and Hellinga (*3*) indicate that in the past, statistical models have been widely used to analyze road crashes in order to explain the relationship between crash involvement and traffic, geometric and environmental factors. However, Chen and Jovanis (*4*) demonstrate certain problems that may arise when using classic statistical analysis on datasets with large dimensions such as an exponential increase in the number of parameters as the number of variables increases and the invalidity of statistical tests as a consequence of sparse data in large contingency tables. This is where data mining comes into play. Data mining can be defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from large amounts of data (5). The use of data mining methods can therefore be particularly useful in the context of large datasets on road accidents.

In this paper, a comparative analysis between high frequency and low frequency accident locations is conducted to determine the discriminating character of the accident characteristics of black spots and black zones. In particular, the data mining technique of association rules is used to obtain a descriptive analysis of the accident data. In contrast with predictive models, the strength of this algorithm lies within the identification of relevant variables that make a strong contribution towards a better understanding of the circumstances in which the accidents have occurred. Hereby, the emphasis will lie on the interpretation of the results, which will be of high importance for improving traffic policies and ensuring traffic safety on the roads.

The paper is organized as follows. First a formal introduction to the technique of association rules is provided. This will be followed by a description of the dataset. Next the results of the empirical study are presented. The paper will be completed with a summary of the conclusions and directions for future research.

ASSOCIATION RULES

Association rules is a data mining technique that can be used to efficiently search for interesting information in large amounts of data. More specifically, the association algorithm produces a set of rules describing underlying patterns in the data by means of the support parameter and the confidence parameter. Informally, the support of an association rule indicates how frequent that rule occurs in the data. The higher the support of the rule, the more prevalent the rule is. Confidence is a measure of the reliability of an association rule. The higher the confidence of the rule, the more confident we are that the rule really uncovers the underlying relationships in the data. It is obvious that we are especially interested in association rules that have a high support and a high confidence.

The concepts behind association rules and suggested algorithms for finding such rules were first introduced by Agrawal, Imielinski & Swami (*6*). They provided the following formal description of this technique:

Let $I = \{i_1, i_2, \ldots, i_k\}$ be a set of literals, called items (accident characteristics). Let D be a set of transactions (accidents), where each transaction T is a set of items such that $T \subseteq I$. Associated with each transaction is a unique identifier, called its *TID*. We say that a transaction T contains X, a set of some items in I, if $X \subseteq T$. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \varnothing$ The rule $X \Rightarrow Y$ holds in the transaction set D with confidence c if c% of transactions in D that contain X also contain Y. The rule $X \Rightarrow Y$ has support s in the transaction set D if s% of transactions in D contain $X \cup$ Y. Given a set of transactions D, the problem of mining association rules is to generate all association rules that have support and confidence greater than the user-specified minimum support *(minsup)* and minimum confidence *(minconf).*

Generating association rules involves looking for so-called *frequent item sets* in the data. Indeed, the support of the rule $X \Rightarrow Y$ equals the frequency of the item set $\{X, Y\}$. Thus by looking for frequent item sets, we can determine the support of each rule (*7*). The problem of discovering association rules can therefore be decomposed into two sub-problems:

1. Generating all item sets that have a support higher than the user-defined minsup. These item sets are called *frequent item set*s.

2. Use this collection of frequent sets to generate the rules that have confidence higher than the user-defined minimum confidence.

DEFINITION 1 Frequency of an item set

s(X, D) represents the frequency of item set X in D, i.e. the fraction of transactions of D that contain X.

DEFINITION 2 Frequent item set

An item set X is called frequent in D, if $s(X, D) \geq \sigma$ *with* σ *the minsup.*

A typical approach (8) to discover all frequent sets X is to use the insight that all subsets of a frequent set must also be frequent. This insight simplifies the discovery of all frequent sets considerably, i.e. first find all frequent sets of size 1 by reading the data once and recording the number of times each item A occurs. Then, form *candidate* sets of size 2 by taking all pairs $\{B, C\}$ of items such that $\{B\}$ and $\{C\}$ both are frequent. The frequency of the candidate sets is again evaluated against the database. Once frequent sets of size 2 are known, candidate sets of size 3 can be formed; these are sets $\{B, C, D\}$ such that $\{B, C\}$, $\{B, D\}$ and $\{C, D\}$ are all frequent. This process is continued until no more candidate sets can be formed. Once all frequent sets are known, finding association rules is easy. Namely, for each frequent set *X* and each *Y*∈ *X* verify whether the rule $X \setminus \{Y\} \Rightarrow Y$ has sufficiently high confidence. The given algorithm has to read the database at most K+1 times, where *K* is the size of the largest frequent set.

The association algorithm generates all rules that have confidence and support higher than minconf and minsup. This implies that all rules with high s(X,Y) will always be generated even if there is no statistically significant dependence between the *X* and *Y* item sets. Therefore, a large subset of the generated rules set will be trivial and a filter is needed to post-process the discovered association rules and to retain only statistically significant accident patterns.

Two properties of association rules can be used to distinguish trivial from non-trivial rules. A first, more formal method (*9*) to assess the dependence between the two item sets in the association rule is lift (L).

DEFINITION 3 Lift

$$L = \frac{s(X \Rightarrow Y)}{s(X) * s(Y)}$$

The nominator s(X⇒Y) measures the observed frequency of the co-occurrence of the items in the antecedent (X) and the consequent (Y) of the rule. The denominator s(X) * s(Y) measures the expected frequency of the co-occurrence of the items in the antecedent and the consequent of the rule under the assumption of conditional independence. The more this ratio differs from 1, the stronger the dependence. Table 1 illustrates the three possible outcomes for the lift value and their associated interpretation for the dependence between the items in the antecedent and consequent of the rule.

TABLE 1 Interpretation of Lift

Outcome	**Interpretation**
$+\infty > L > 1$	Positive interdependence effects between *X* and *Y*
$L = 1$	Conditional independence between *X* and Y
$0 < L < 1$	Negative interdependence effects between *X* and *Y*

A second method to discern trivial from non-trivial rules is looking at the statistical rule significance (*10*).

DEFINITION 4 Statistical Rule Significance

The statistical significance of a rule is the validity of a rule, based on the influence of statistical dependency between the rule body (antecedent) and the rule head (consequent).

The statistical significance (T) of a rule is the validity of a rule, based on the influence of statistical dependency between the rule body and the rule head.

T is determined using the χ^2- test for statistical independence:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

If the χ^2 value < critical χ^2 value (p=0.5, 3 degrees of freedom), there is statistical independency between the rule body and the rule head and the statistical significance (T) will be neutral. If the χ^2 value > critical χ^2 value, there is statistical dependency between the rule body and the rule head. Depending on the relationship between the observed *(Oij)* and the expected *(Eij)* frequencies, the algorithm determines whether the statistical significance (T) is negative or positive.

Table 2 gives an illustration for the possible outcomes of the statistical rule significance test (T) and indicates its relation with the lift value of the rule.

- TABLE 2 Interpretation of Statistical Rule Significance

Outcome	***Interpretation***
T <0	- Item set X has a negative influence on the occurrences of item set Y - Given item set X, item set Y occurs less frequently than expected - Lift between 0 and 1 - Valid rule (T = -)
T is neutral	- Lift =1: X and Y are statistically independent → Rule gives no extra information - Lift ≠1: rule has failed the χ^2- test → Rule is not valid
T >0	- Item set X has a positive influence on the occurrences of item set Y - Given item set X, Y occurs more frequently than expected - Lift >1 - Valid rule (T = +)

DATA

This study is based on a large data set of traffic accidents obtained from the National Institute of Statistics (NIS) for the region of Flanders (Belgium) for the year 1999. Given the limitation that this one year period is not long enough to limit random fluctuations, the period under study does limit changes in road and traffic conditions. More specifically, the data are obtained from the Belgian "Analysis Form for Traffic Accidents" that should be filled out by a police officer for each traffic accident that occurs with injured or deadly wounded casualties on a public road in Belgium. In total, 34.353 traffic accident records are available for analysis.

The traffic accident data contain a rich source of information on the different circumstances in which the accidents have occurred: course of the accident (type of collision, road users, injuries, …), traffic conditions (maximum speed, priority regulation, …), environmental conditions (weather, light conditions, time of the accident, …), road conditions (road surface, obstacles, …), human conditions (fatigue, alcohol, …) and geographical conditions (location, physical characteristics, …). On average, 45 attributes are available for each accident in the data set.

An initial analysis on the dataset indicated that the traffic accident data are highly skewed. This means that some of the attributes will have an almost constant value for each of the accidents in the database. For example, 80% of the accidents in the dataset occurred under normal weather conditions. As was explained earlier, this will have no effect on the validity of the results since the association algorithm produces the lift value that corrects the importance of each rule by taking the frequency of the attributes in the dataset into account.

EMPIRICAL STUDY

We distinguish three steps in the mining process: a preprocessing step in which the available data is prepared for the optimal use of the mining technique, a mining step for generating the association rules and a post-processing step for identifying the most interesting association rules.

Preprocessing the dataset

To discern high frequency accident locations from low frequency accident locations, each accident needs to be linked with a location parameter that corresponds to a unique geographical location. Therefore, the accidents that occurred at district and province roads are located by means of the road identification number and the kilometer mark. The accidents that took place at a non-numbered road are located using the street name and the name of the city in which the accident occurred.

Next, two different data sets were selected to explore association relationships between traffic accident attributes. Since our prime interest lies in the profiling and understanding of black spots and black zones, only the traffic accidents that occurred at a high frequency accident location were selected for the first analysis. This allows us to give a descriptive analysis of frequently occurring accident patterns on highly concentrated accident locations. To identify these locations, a criterion of minimum five accidents per location was used. This resulted in a total of 3.368 traffic accident records that were included in the first analysis. This number of accidents corresponds with the fact that in Flanders 15% of all the traffic accidents occur on so-called dangerous spots (*11*). The second association analysis is carried out on the remaining low frequency accident locations, including 30.985 accidents. By comparing the results from these two analyses, we can determine the discriminating character of the accident characteristics of high frequency accident locations.

Furthermore, in the present data set, some attributes have a continuous character. Discretization of these continuous attributes is necessary, since generating association rules requires a data set for which all attributes are discrete. Therefore, the observations for these variables are divided into different intervals by grouping them into partitions. For example, six new attributes were created from the continuous variable 'time of accident': morning rush hour (7h-9h), morning (10h-12h), afternoon (13h-15h), evening rush hour (16h-18h), evening (19h-21) and night (22h-6h). The intervals for this variable were created on the basis of expert knowledge of traffic rush hours in Belgium. For those variables where no domain knowledge for grouping the attributes could be found, we used the *Equal Frequency Binning* discretization method to generate intervals containing an equal number of observations (*12*).

Finally, attributes with nominal values had to be transformed into attributes with binary attribute values. This means that dummy variables had to be created by associating a binary attribute to each nominal attribute value of the original attributes.

Generating Association Rules

A minimum support value of 5 percent was chosen for the analysis. This means that no item or set of items will be considered frequent for the first analysis if it does not appear in at least 165 traffic accidents. Obviously the same threshold will be used for the second analysis since the main purpose of this research involves a comparative analysis between the rules sets of the high frequency accident locations and the low frequency accident locations. It could be argued that the choices for the values of these parameters are rather subjective. This is partially true, however a trial and error experiment indicated that setting the minimum support too low, leads to exponential growth of the number of items in the frequent item sets. Accordingly, the number of rules that will be generated will cause further research on these results to be impossible due to computer memory limitations. In contrast, by choosing a support parameter that is too high, the algorithm will only be capable of generating trivial rules. Analogously, the minimum confidence value was set at 30 percent. This means that a rule is considered reliable when the consequent of the rule occurs at least one out of three times that the antecedent appears. By choosing different confidence values, a trial and error experiment showed that this parameter value gives rather stable results concerning the amount of rules generated by the algorithm.

From the high frequency accident locations, with a minsup = 5 percent and minconf = 30 percent, the algorithm obtained 187.829 frequent item sets of maximum size 4 for which 598.584 association rules could be generated. Although these results relate to a relatively small number of accident records, they are quite reasonable since an average of 40 items is available per accident, allowing the algorithm to generate multiple combinations of size 4 item sets. With the same parameters the second analysis resulted for the low frequency accident locations in 183.730 frequent item sets of maximum size 4 for which 575.974 association rules could be generated. These rules are further processed to select the most interesting rules.

Post-processing the Association Rules Set

The purpose of post-processing the association rules set is to identify the subset of interesting (i.e., non-trivial) rules in a generated set of association rules. Selecting the rules with a positive or a negative statistical significance from the association rules set narrowed down the results to 14.690 association rules for the high frequency accident locations and 77.282 association rules for the low frequency accident locations.

However, after ranking the association rules on their lift value and removing the non-significant rules from the rules set, one important problem still remains. The discovered accident patterns for the high frequency accident locations will give a description of the frequently occurring accident circumstances, but they may also be characteristic for the accidents that occur at low frequency accident locations as they represent the necessary but not the sufficient condition for the membership of a high frequency accident location. Therefore, we use the interestingness measure to limit the association rules to only the discriminating or useful ones (*13*).

DEFINITION 5 Interest (I)

$$I = \frac{S_h - S_l}{\max\{S_l, S_h\}}$$

This interestingness measure is based on the deviation of the characteristic rules discovered for the accidents that occurred on high frequency accident locations from the accidents that occurred on low frequency accident locations. The nominator S_h -S_l measures the difference in support for the rules in the high frequency accident data set (S_h) and the low frequency accident data set (S_l). The expression max $\{S_l, S_h\}$ is called the normalizing factor as it normalizes the interestingness measure onto the scale [-1, 1]. Since in this research we are mainly interested in profiling the high frequency accident locations, we will pay special attention to the rules with a positive interest value, i.e. approximating '1'.

RESULTS

As stated earlier, the emphasis in this study lies on the identification and profiling of frequently occurring accident patterns at high frequency accident locations and the degree in which these accident characteristics are discriminating between high frequency and low frequency accident locations. Selecting the association rules that appear in both the high frequency accident rules set and the low frequency accident rules set results in 3.670 statistically significant association rules. These can be further post-processed by means of the interestingness measure.

When ranking the association rules on their interest value, figure 1 shows, for the 50 most discriminating rules, that a high interestingness value does not inseparably correlates with a strong lift value.

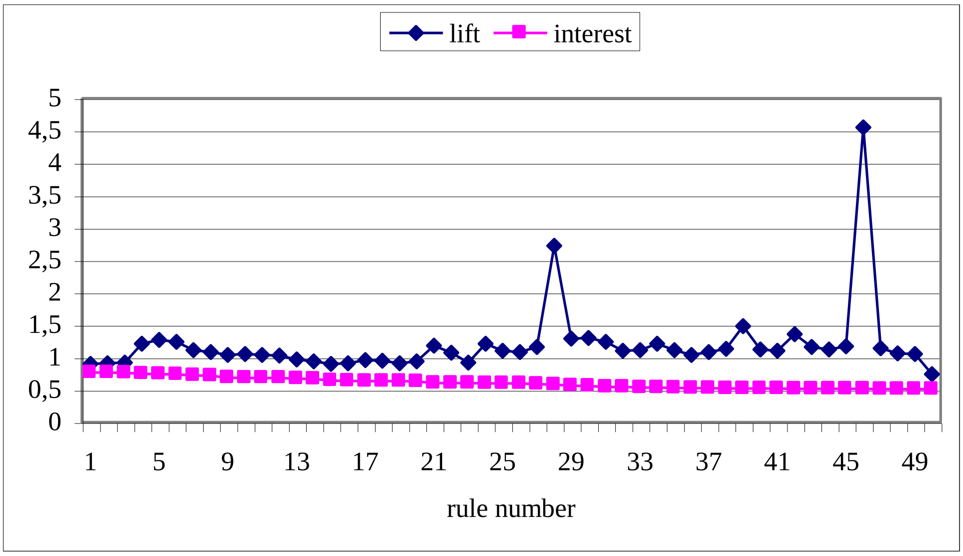

FIGURE 1 Association rules ranked on descending interest value

Note that we do not use the term 'high' for the lift value, since a very small lift value, i.e. considerably differing from 1, also indicates a strong (negative) dependency between the rule body and the rule head. These results show that accident characteristics that have the most discriminating power to identify black spots and black zones are not necessarily the most interesting rules according to their lift values.

Accordingly, when ranking the association rules on the lift value, figure 2 and figure 3 indicate that although the association rules identify frequently occurring patterns that are descriptive for the occurrence of accidents on high frequency accident locations, they are not necessarily discriminating between the profile of high frequency accident locations and low frequency accident locations.

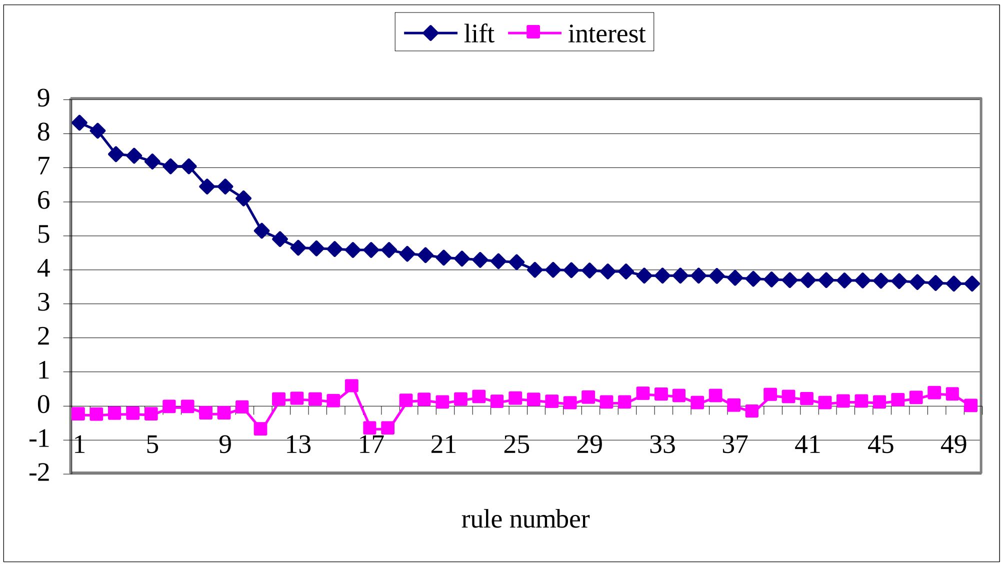

FIGURE 2 Association rules ranked on descending lift value

When looking at the association rules with the highest lift values, figure 2 shows that the majority of these rules have a small interestingness value. Also for the very small lift values, figure 3 indicates that the strong dependencies between different accident characteristics do not always correspond with high interestingness values. In general, these rules with a strong lift value and a low interestingness value are characteristic for both high frequency accident locations and low frequency accident locations.

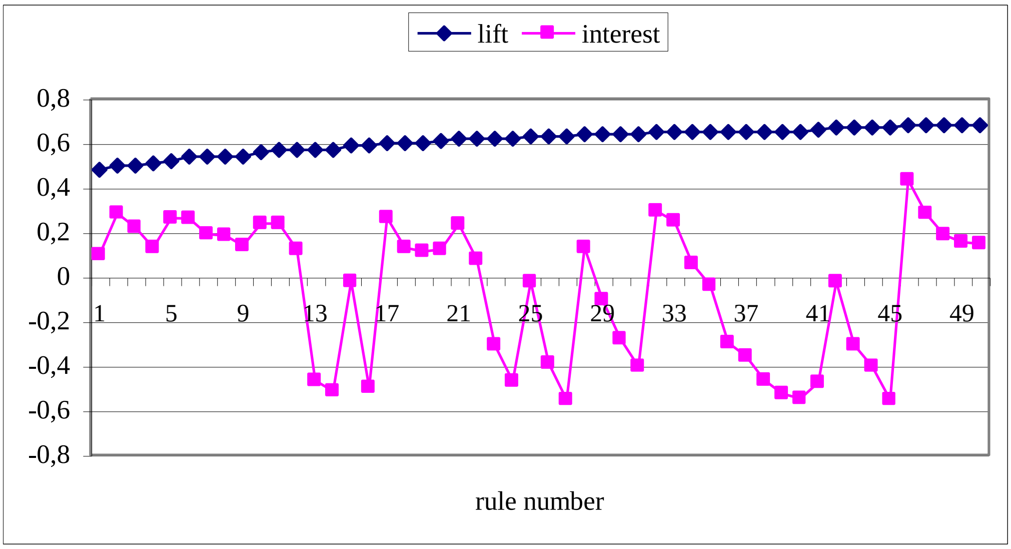

FIGURE 3 Association rules ranked on ascending lift value

When looking at the interpretation of the association rules, the upper part of table 3 shows that the rules with the 10 highest interestingness values mostly relate to geographical characteristics of the accidents. This means that the most discriminating characteristics between high frequency accident locations and low frequency accident locations are related to infrastructure or location features. For example, when an accident occurs on a roadway with separated lanes (by means of a guard-rail or a roadside), it will less frequently than expected take place outside a crossroad. Additionally, an accident occurring on this type of roadway will have a higher probability than expected of occurring outside the inner city. Since the accidents analysed in this research all occurred in the region of Flanders, these results seems quite reasonable since in this part of Belgium most roadways with separated lanes are located outside the inner city. In total 46,4 percent of all accidents that occur at high frequency accident locations take place on a roadway with separated lanes outside the inner city. In comparison with the low frequency accident locations, where only 12,34 percent of the accidents can be attributed to this sort of location, these roadways with separated lanes outside the inner city are very characteristic for the occurrence of black spots and black zones. Although these results seem quite reasonable, they do indicate that roadways with separated lanes outside the inner city are an important problem for traffic safety. Therefore, further research on the cause of the unsafe character of these roads will be necessary. A high traffic intensity would seem the logical explanation for the high number of accidents that occur on these roads, but another possible explanation could also be the infrastructure of these roads. Depending on the results, government could consider restructuring these roads or changing their traffic regulation.

Furthermore, when an accident takes place on a roadway with separated lanes, the road user will more frequently than expected be of the age 30 until 45. Finally, a characteristic pattern for high frequency accident locations is the involvement of at least one passenger car among the road users when the accident occurs on a roadway with separated lanes. Accordingly to the previous results, this kind of accident accounts for 47,86 percent of all accidents on high frequency accident locations, whereas the same accident circumstances only occur in 15,02 percent of the low frequency accident locations.

The middle part of table 3 gives the 10 most important rules for the high frequency accident locations based on the descending lift values. The negative interest values of these rules indicate that the accident circumstances described in these patterns have a higher occurrence on the low frequency accident locations than on the high frequency accident locations, although the differences in support values are very small. Furthermore, the results show that the strongest positive dependencies between accident characteristics are not as strongly location related. Most of these association rules refer to the number of persons involved in the accident and the number of casualties following the accident. This can be explained by the very small interestingness values of the rules, referring to the occurrence of the accident characteristics at both low frequency accident locations and high frequency accident locations. As a result, these rules will identify patterns that will be less geographical or location related and more human or vehicle related. Two exceptions can be made. The association rules concerning cyclists do refer to frequently occurring infrastructure characteristics. These patterns state that when an accident occurs with a cyclist who is riding on a cycle track that is marked on the roadway, the type of this cycle track will more frequently than expected be a one-way track and vice versa. This kind of accident accounts for 7 percent of all accidents on high frequency accident locations but also for 5,78 percent of all accidents on low frequency accident locations. Therefore, these accident characteristics are not very discriminating between high and low frequency accident locations, but they do however identify an important problem in the traffic safety of cyclists in general.

TABLE 3 Rules for High and Low Frequency Accident Locations

BODY	HEAD	I	T	L	Sh	Sl	CONF
10 Rules with highest interest value							
[Belgian road user]+[roadway separated lanes]	=> [outside crossroad]	0,76	-	0,90	36,46	8,51	70,62
[road user in normal condition]+[roadway separated lanes]	=> [outside crossroad]	0,76	-	0,91	37,23	8,77	70,93
[roadway separated lanes]	=> [outside crossroad]	0,76	-	0,92	39,81	9,54	71,60

[daylight]+[roadway separated lanes]	=>	[outside inner city]	0,75	+	1,21	30,87	7,81	81,91
[Belgian road user]+[roadway separated lanes]	=>	[outside inner city]	0,74	+	1,27	44,12	11,40	85,45
[roadway separated lanes]	=>	[outside inner city]	0,73	+	1,24	46,40	12,34	83,45
[road user in normal condition]+[roadway separated lanes]	=>	[age 30-45]	0,72	+	1,11	31,79	8,91	60,58
[roadway separated lanes]	=>	[age 30-45]	0,72	+	1,08	32,98	9,31	59,32
[roadway separated lanes]	=>	[passenger car]	0,69	+	1,04	47,86	15,02	86,07
[road user in normal condition]+[roadway separated lanes]	=>	[Belgian road user]	0,68	+	1,05	49,02	12,52	93,38
10 Rules with highest lift value								
[0 passengers]+[1 seriously injured]	=>	[0 lightly injured]	-0,31	+	8,28	7,33	10,56	85,76
[0 deadly injured]+[0 passengers]+ [0 lightly injured]	=>	[1 seriously injured]	-0,31	+	8,05	7,21	10,49	96,05
[0 deadly injured]+[0 deaths after serious injuries] +[0 seriously injured]	=>	[1 seriously injured]	-0,29	+	7,36	7,69	10,79	87,80
[0 deadly injured]+[0 lightly injured]	=>	[1 seriously injured]	-0,29	+	7,31	7,69	10,80	87,21
[0 passengers]+[0 lightly injured]	=>	[1 seriously injured]	-0,31	+	7,14	7,33	10,56	85,17
[cycle track marked on roadway]	=>	[one-way cycle track]	-0,08	+	7,00	5,78	6,31	82,98
[one-way cycle track]	=>	[cycle track marked on roadway]	-0,08	+	7,00	5,78	6,31	48,87
[1 seriously injured]	=>	[0 lightly injured]	-0,28	+	6,41	7,92	11,02	66,42
[0 lightly injured]	=>	[1 seriously injured]	-0,28	+	6,41	7,92	11,02	76,50
[outside inner city]+[1 seriously injured]	=>	[0 lightly injured]	-0,10	+	6,06	5,81	6,49	62,82
10 Rules with smallest lift value								
[wet road surface]	=>	[normal weather conditions]	0,10	-	0,48	11,19	10,09	38,47
[collision with obstacle outside roadway]	=>	[continuing driving direction]	0,28	-	0,50	7,77	5,56	38,87
[loss control steering wheel]	=>	[2 road users]	0,22	-	0,50	7,92	6,17	32,40
[Belgian road user]+[2 lightly injured]	=>	[0 passengers]	0,13	-	0,51	5,84	5,08	36,21
[Belgian road user]+[loss control steering wheel]	=>	[2 road users]	0,26	-	0,52	7,42	5,47	34,01
[passenger car]+[loss control steering wheel]	=>	[2 road users]	0,26	-	0,54	6,91	5,10	35,30
[male road user]+[loss control steering wheel]	=>	[2 road users]	0,19	-	0,54	6,97	5,63	34,92
[no alcohol]+[loss control steering wheel]	=>	[2 road users]	0,19	-	0,54	6,35	5,17	34,97
[Belgian road user]+[0 seriously injured+[2 lightly injured]	=>	[0 passengers]	0,14	-	0,54	5,84	5,03	38,33
[road user normal condition]+[loss control steering wheel]	=>	[2 road users]	0,24	-	0,56	7,74	5,89	36,20

The results for the 10 strongest negative dependency rules are shown, based on the ascending lift values, in the lower part of table 3. The association rule with the smallest lift value indicates that when an accident occurs on a wet road surface, the weather will less frequently than expected be normal. Obviously this is a very strong dependency that is valid for most accidents, whether they occur on high frequency accident locations or not. The second most important rule of this table indicates that when a road user crashes into an obstacle outside the roadway, he was less frequently than expected continuing his driving direction. Furthermore, when losing control over the steering wheel, there is a smaller probability than expected that two road users be involved in the accident. Corresponding with the results for the positive rule dependencies, these patterns relate mainly to human characteristics and less to location related circumstances. However, a small difference in interpretation can be noted since the negative dependency rules mainly describe behavioural aspects of traffic accidents indicating different accident types. It seems that losing control over the steering wheel and crashing into an obstacle outside the roadway are frequently occurring accident patterns, each with different accident circumstances. Accordingly, these association rules are more descriptive for the occurrence of accidents in general and they are less discriminating between high frequency accident locations and low frequency accident locations.
When looking at the results of table 3, a final remark concerning the interest value of the rules should be made. One can see that the support values for the patterns in the low accident locations remain quite stable over the tree parts of the table. However, the support values for the rules concerning the high frequency accident locations are considerably high for the upper part of the table and considerably smaller for the middle and lower part of the table. Therefore, the increase in the interest value of the rules and consequently the increase in the discriminating character of the rules are mainly related to the strong occurrence of the accident circumstances in the high frequency accident locations and not as much by the weak occurrence of these patterns in the low frequency accident locations.

Special attention should also be given to the association rules that do appear in the rules set of the high frequency accident locations but not of the low frequency accident locations. Table 4 shows the results for these association rules with the 10 highest lift values. The rules are mainly related with the number of casualties of the accident but also with one specific accident type: collision with an obstacle outside the roadway. When this kind of collision occurs on a roadway with separated lanes outside a crossroad or outside the inner city, the obstacle will more frequently than expected be a crash barrier (iron or concrete).

TABLE 4 Rules for High Frequency Accident Locations

BODY		HEAD	T	L	S_h	CONF
[0 deaths after serious injuries]+[0 passengers] +[0 lightly injured]	=>	[1 seriously injured]	+	7,19	7,33	85,76
[normal weather]+[1 seriously injured]	=>	[0 lightly injured]	+	6,73	6,65	69,78
[normal physical condition]+[0 deaths after serious injuries] +[0 lightly injured]	=>	[1 seriously injured]	+	6,71	7,30	80,13
[normal physical condition]+[0 lightly injured]	=>	[1 seriously injured]	+	6,67	7,30	79,61
[roadway with separated lanes]+[0 seriously injured] +[collision with obstacle outside roadway]	=>	[crash barrier]	+	6,51	5,52	52,39
[roadway with separated lanes]+[outside crossroad] +[collision with obstacle outside roadway]	=>	[crash barrier]	+	6,34	6,74	51,01
[roadway with separated lanes]+[outside inner city] +[collision with obstacle outside roadway]	=>	[crash barrier]	+	6,31	6,77	50,78
[roadway with separated lanes]+[0 deaths] +[collision with obstacle outside roadway]	=>	[crash barrier]	+	6,21	6,68	50,00
[outside inner city]+[0 seriously injured] +[collision with obstacle outside roadway]	=>	[crash barrier]	+	6,05	6,56	48,68
[roadway with separated lanes] +[collision with obstacle outside roadway]	=>	[crash barrier]	+	6,03	6,80	48,52

In the most favourable situation, these rules are not characteristic for the low frequency accident locations (S_l=0, Interest=1) and consequently these patterns optimally discriminate between high frequency accident locations and low frequency accident locations. However, in the most pessimistic situation these patterns do exist for the low frequency accident locations but they do not appear in the association rules set due to the value of the minimum support parameter (5 percent). Moreover, the choice of the minimum confidence parameter (30 percent) could inhibit the algorithm from generating a number of rules although the frequent item sets exceed the minimum support parameter. Consequently, although these association rules could give some valuable information on the occurrence of traffic accidents in general, no conclusion can be made towards the discriminating character of these rules between high an low frequency accident locations. Furthermore, the support value (or the occurrence) of these accident circumstances is relatively low for the high frequency accident locations and therefore we will no further go in to detail about the interpretation of these rules.

CONCLUSIONS AND FUTURE RESEARCH

In this paper, the technique of association rules was used on a large dataset of traffic accidents to profile black spots in terms of accident related data and location characteristics. The analysis showed that by generating association rules the identification of accident circumstances that frequently occur together is facilitated. This leads to a strong contribution towards a better understanding of the occurrence of traffic accidents. However, association rules do describe the co-occurrence of accident circumstances but they do not give any explanation about the causality of these accident patterns. Therefore, their role is to give direction to more profound research since the use of some additional techniques or expert knowledge will be required to identify the most important causes of these accident patterns, allowing governments to better adapt their traffic policies to the different kind of accident circumstances. Furthermore, the results indicate that the use of the association algorithm not only allows to give a descriptive analysis of accident patterns on high frequency accident locations, it also creates the possibility to find the accident characteristics that are discriminating between high frequency accident locations and low frequency accident locations. The most important results indicate that although human and behavioural characteristics play an important role in the occurrence of all traffic accidents, the main difference in accident patterns between high frequency accident locations and low frequency accident locations can be found in infrastructure or location related circumstances. In conclusion, this analysis shows that a special traffic policy towards black spots and black zones should be considered, since these high frequency accident locations are characterized by specific accident circumstances, which require different measures to improve the traffic safety.

Although the analysis carried out in this paper revealed several interesting rules, which, in turn, provide valuable input for purposive government traffic safety actions, several issues remain for future research. First, the skewed character of the accident data limits the amount of information contained in the dataset and will therefore restrict the number of circumstances that will appear in the results. Moreover, the choice for the minimum support and minimum confidence parameter can prevent the association algorithm from generating rules on the less frequent accident conditions. Secondly, the inclusion of domain knowledge (e.g. traffic intensities, a priori infrastructure distributions) in the association algorithm would improve the mining capability of this data mining technique and would facilitate the post-processing of the association rules set to discover the most interesting accident patterns. Finally, considering the large number of attributes in the traffic accident dataset, it seems interesting to explore the potential of techniques that generate rules with long patterns to uncover more complex associations in traffic accidents.

REFERENCES

(1) Belgian Institute for Traffic Safety (BIVV) and National Institute for Statistics, *Year Report on Traffic Safety 2000* (CD-ROM), BIVV v.z.w., Brussels.

(2) Kononov, J. and Janson B. Diagnostic Methodology for the detection of safety problems at intersections. In *Proceedings of the Transportation Research Board* (CD-ROM), Washington D.C., USA, January 13-17, 2002.

(3)Lee, C., Saccomanno, F. and B. Hellinga. Analysis of Crash Precursors on Instrumented Freeways, In *Proceedings of the Transportation Research Board* (CD-ROM), Washington D.C., USA, January 13-17, 2002.

(4) Chen, W. and P. Jovanis. Method for identifying factors contributing to driver-injury severity in traffic crashes. In *Transportation Research Record 1717*, TRB, National Research Council, Washington, D.C., 2002, pp. 1-9.

(5) Frawley, W., Piatetsky-Shapiro, G., and C. Matheus. Knowledge discovery in databases: an overview. *Knowledge Discovery in Databases.* AAAI Press/ MIT Press, Menlo Park, California, USA, 1991 pp. 1-27.

(6) Agrawal, R., Imielinski, T. and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of ACM SIGMOD Conference on Management of Data*, Washington D.C., USA, May 26-28, 1993, pp. 207-216.

(7) Mannila, H. Methods and problems in data mining. In *Proceedings of the International Conference on Database Theory*, Delphi, Greece, January 8-10, 1997, pp. 41-45.

(8) Agrawal, R., Mannila, H., Srikant, R. et al. Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining.* AAAI Press, Menlo Park, California, USA, 1996, pp. 307-328.

(9) Brin, S., Motwani, R. and C. Silverstein. Beyond market baskets: generalizing association rules to correlations. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, Tucson, Arizona, USA, May 13-15, 1997, pp. 265-276.

(10) Silverstein, C., Brin, S. and R. Motwani. Beyond market baskets: generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, Vol. 2, No.1, 1998, pp. 39-68.

(11) Ministry of Flemish Government. Design mobility plan Flanders, Belgium, 2001, http://viwc.lin.vlaanderen.be/mobiliteit/

(12) Holte, R.C. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, Vol. 11, 1993, pp. 63-90.

(13) S. S. Anand, D. A. Bell, J. G. Hughes, A. Patrick. Tackling the cross sales problem using data mining. In *Proceedings of the 1st International Conference On Knowledge Discovery and Data Mining*, 1997.