Research on Traffic-Related Events Based on Big Data

Authors: He Qun, He Qun

Date: 2024-03-19T00:00:00+00:00

Abstract

Abstract: This paper presents an analysis and solution approach based on big data resource mining for traffic incidents among motor vehicles, non-motor vehicles, and pedestrians in ordinary road traffic. The conditions and consequences of such incidents are organized into a relational table, from which rules are extracted using relational database routines based on association theory. Analysis of these rules yields methods for mitigating incidents among the three entity types. Data sources comprise quasi-big-data platforms from various traffic authorities, enhancing data depth and breadth. Algorithmically, a method for converting data itemsets to relational tables is proposed, along with an associability degree algorithm for table attributes. Attribute associability serves as the criterion for optimal test set selection, enabling primary optimization at the attribute level. This departs from traditional Apriori algorithms that treat all items in data itemsets equally and rely on frequent pattern judgment, thereby reducing computational intensity. The principle difference lies in narrowing search scope via attribute associability rather than redundant enumeration and pruning; the processing difference lies in utilizing database system routines directly rather than time-consuming programming with frequent database I/O.

Full Text

Preamble

Research on Traffic-Related Events Based on Big Data. Current traffic regulations categorize road users into three classes: motor vehicles, non-motor vehicles, and pedestrians, each governed by specific rules. However, extracting actionable knowledge from frequent incidents to guide the behavior of these three entities and prevent traffic events represents a significant research challenge. This constitutes a data mining problem [1], requiring both comprehensive traffic incident data and appropriate mathematical frameworks for information processing. With the advancement of the big data industry [2][3], large-scale traffic incident

datasets have become available; association theory [4] provides one viable mathematical approach for analyzing such data. This paper proposes an analytical method for traffic-related event information that integrates big data with association algorithms through SQL routines [5][6]. By analyzing extensive incident data, we derive regular patterns and offer recommendations regarding critical situations requiring attention, behavioral habits needing modification, and traffic regulations warranting revision.

2 Traffic Association Problems

Common traffic association problems among the three entity types include: (1) Insufficient prediction [7], where drivers fail to adequately anticipate the movements of entities on their right; (2) Vehicle loss of control [8-9], caused by driver operational errors; (3) Driving blind spots [10], which occur when turning right alongside other moving entities due to insufficient clearance; (4) Night driving [11], where reduced visibility contributes to incidents; (5) Driving retaliation [12], involving deliberate aggressive actions; (6) Tracking while moving [13], resulting from driver distraction; (7) Sudden obstacles, such as thrown objects or vehicles stopping abruptly and opening right-side doors; and (8) Snatching incidents, which exploit speed differentials between entity types for dynamic theft.

3 Data Sources

The 2015 National Outline for Promoting Big Data Development systematically deployed national big data initiatives, proposing the formation of cross-departmental data resource sharing frameworks by 2017. Current departmental data platforms can be considered subsets of broader big data ecosystems [14][15]. The role of traffic safety big data platforms is to rationally collect, manage, and process valuable information from vast datasets to obtain safety-relevant insights.

Based on the 4V characteristics of big data [2], we can effectively understand and apply these platforms: **Volume** and **Value** require targeted data acquisition from large platforms, meaning we must historically and comprehensively collect data related to the eight categories of problems described above. **Velocity** and **Variety** present challenges as high-speed data acquisition involves numerous format types, necessitating conversion of collected data into structured traffic event data sources. Various unstructured data must be transformed into structured formats according to algorithmic requirements. For traffic events, unstructured data such as scene records, surveillance footage, and liability certifications must be converted into formatted symbolic data [14] to form relational tables—the fundamental operational objects of relational databases.

Our primary data source consists of current traffic department data platforms [15], supplemented by traffic incident information from departmental official

websites, online event publications, analyses, and public discussions, collectively forming a comprehensive dataset for mining.

Incidents among the three entities represent an association problem. The mathematical foundation for data association mining is the Apriori algorithm for association rule mining [4][16], which identifies relationships between items in a dataset. Its core principle implements two key definitions—support and confidence—through two stages: frequent itemset generation and rule testing.

Definition 1 (Support). Support is the ratio of the number of test sets containing X to the total sample size D: supp(X)=occur(X)/count(D)=P(X). For example, P(A B) represents the probability of both A and B occurring.

Definition 2 (Confidence). Confidence is the ratio of the number of test sets containing both X and Y to the number containing X: conf(X→Y)=supp(X Y)/supp(X)=P(Y|X). For example, P(B|A)=P(AB)/P(A) represents the probability of B occurring given that A has occurred.

The Apriori algorithm employs an iterative, layer-by-layer search method: first identifying frequent itemsets that satisfy support thresholds, then deriving strong rules from these itemsets that meet both minimum support and minimum confidence criteria. The primary bottleneck lies in the incremental search for candidate itemsets, while another significant time-consuming factor involves extensive I/O operations from repeated database scanning [16]. Therefore, optimizing search methods and database interactions is critical for algorithmic improvement.

To this end, we propose an improved algorithm based on relational table processing: (1) Define each item type in the original dataset (Itemset) as table attribute names, with each transaction identifier (Tid) as a record. Use attribute values {1,0} to represent relationships between items, thereby equivalently converting the dataset into a relational table. For instance, the classic dataset in Table 1 is transformed into the relational table shown in Table 2; (2) Extend attribute values in such relational tables to multiple values for broader applicability; (3) For each attribute, calculate its relatedness degree (see Definition 3) and select superior attributes into the test area based on relatedness strength; (4) Use database-supported routines to calculate support and confidence between attributes.

Definition 3 (Relatedness Degree). Let attribute X have values $\{x_1, x_2, \cdots, x\}$. Relatedness degree is a test calculation that identifies the most frequent attribute value and computes the ratio of its count to the total number of distinct attribute value categories: rltd(X)=Maxcount(x)/Ccount(k)=Maxcount(x)/j (k=1,2,⋯,j). Here, Maxcount(x) finds the highest frequency of any value in attribute X, while Ccount(k) determines the number of distinct value categories (j). A larger rltd(I) value indicates higher relatedness potential of attribute X.

For example, testing relatedness degrees for attributes in Table 2 yields: rltd(A)=Maxcount(A=1)/Ccount({0,1})=count(1)/2=2/2=1; rltd(B)=Maxcount(B=1)/Ccount({0,1})=count

similarly, rltd(C)=1.7; rltd(D)=0.5; rltd(E)=1.7. Thus, B, C, and E are preferred attributes, A is secondary, and D is the weakest. Consequently, only B, C, and E require confidence testing:

Sel @s=count(B)/count(*) from Tab2 where (B=1)and(C=1)and(E=1)

Here, credibility @s=2/4=0.5=50%, matching the conclusion of the classical algorithm.

Definitions 1 and 2 provide inter-item association testing methods, while Definition 3 measures single-attribute relatedness from an attribute perspective. The algorithm uses relatedness degree to constrain data selection scope and performs selection operations on data tables to derive conclusions. The fundamental difference from traditional Apriori lies in using relatedness degree to narrow the search space rather than complex enumeration and pruning; processing directly utilizes database system routines [5][6] instead of programmed database I/O interactions. This approach transforms the traditional Apriori algorithm' s computational intensity of simultaneously finding all frequent itemsets into an attribute-wise data itemset reduction that leverages underlying database routines for sorting and searching, thereby improving computational speed. The algorithm proceeds as follows: (1) Build a relational table from the itemset; (2) Calculate each attribute' s relatedness degree per Definition 3; (3) Select attributes into the test area based on relatedness strength; (4) Perform selection query operations on the test area per Definitions 1 and 2 to obtain support and confidence for various attribute value combinations; (5) Output rules satisfying the support and confidence thresholds.

5 Mining Process

Effective data mining requires both targeted focus and sufficient depth and breadth. Depth refers to the historical time span covered, while breadth concerns the geographic scope. This study analyzes 250 typical traffic incidents involving the three entity types across multiple provinces in recent years. Due to space constraints while ensuring comprehensive illustration, we consolidate and compress incident data types as shown in Table 3 , where each record sequence number N represents n similar incidents.

We define six key attributes: Event Type (E), Road Level (L), Subject State (M), Object State (O), Involved Relationship (R), and Loss Conclusion (C).

Attribute values are specified as follows: - **Event Type (E):** insufficient prediction (a), loss of control (b), resolution obstacle (c), sudden incident (d), where a, b, c, and d respectively encompass problems (1) & (6), (2) & (5), (3) & (4), and (7) & (8) from our earlier classification. - **Road Level (L):** secondary isolation (a), primary isolation (b), no isolation (c). - **Subject State (M):** sharp left turn (a), sharp right turn (b), other (c). - **Object State (O):** facing (a), back-facing (b), mixed (c). - **Involved Relationship (R):** motor-pedestrian (a), motor-non-motor (b), non-motor-pedestrian (c), motor-non-motor-pedestrian (d). -

Loss Conclusion (C): severe (a), heavy (b), moderate (c), general (d).

Following algorithm step 1, we synthesize relational tables for each incident based on scene records and related materials from Table 3, producing Table 4 (named Db). Applying algorithm steps 2 and 3 to the E=a event subset, we calculate relatedness degrees per Definition 3:

rltd(R)=Maxcount(R.x)/Ccount(R)=count(a)/3=6/3=2
rltd(C)=Maxcount(C.x)/Ccount(C)=count(b)/4=4/4=1
rltd(L)=Maxcount(L.x)/Ccount(L)=count(b)/4=5/4=1.25
rltd(M)=Maxcount(M.x)/Ccount(M)=count(b)/2=7/2=3.5
rltd(O)=Maxcount(O.x)/Ccount(O)=count(b)/2=7/2=3.5

Attributes M and O exhibit the strongest relatedness and are selected for the test area. Per algorithm step 4 and Definitions 1-2, we calculate support and confidence for attribute values (M=b, O=b):

```
declare @n,@S,@D ;
Sel @n=count(n) from Db where (E=a);
Sel @S=count(n)/@n from Db where ((E=a) and (M=b) and (O=b));
Sel @D=@S*@n/count(n) from Db where ((E=a) and (O=b));
```

The results show @S (support)=62/75=0.8 and @D (confidence)=(62/75)/(62/75)=1.

Following algorithm step 5, we output these support and confidence values as equation (1) in Table 4 (right). Iterating steps 3-5 yields other attribute combinations with maximum support and confidence. Similarly, we obtain typical support and confidence values for events where E=b, c, and d, as presented in Table 4 (right).

6 Rule Analysis

For insufficient prediction events (E=a): Equation (1) in Table 4 (E=a) shows S=0.8 and D=1, indicating that incidents where the subject entity suddenly moves rightward while the object entity faces away account for 80% of this category with 100% confidence. Equations (2) and (3) further reveal that support and confidence levels for heavy losses and motor-pedestrian involvement reach (60%, 70%) and (62%, 58%) respectively.

For loss of control and driving retaliation events (E=b): Equation (1) demonstrates that motor-pedestrian incidents constitute 60% of this category with 100% confidence, while equation (2) shows support and confidence for heavy-or-greater losses reach 50% and 100% respectively.

For resolution obstacle events (E=c): Equation (1) indicates that incidents featuring sudden rightward movement by the subject, back-facing orientation of the object, and heavy losses achieve support and confidence of 70% and 100% respectively.

For sudden incident events (E=d): Equation (1) shows that incidents with

sudden rightward movement and back-facing objects represent 40% of this category with 100% confidence. Equations (2) and (3) further demonstrate support and confidence for general losses and motor-pedestrian involvement of (50%, 80%) and (30%, 50%) respectively.

7 Analysis Conclusions

The analysis demonstrates that across various traffic events on ordinary roads, incidents characterized by sudden rightward movement of the subject entity and back-facing orientation of the object entity resulting in heavy losses exhibit the highest support and confidence levels. These incidents primarily involve motor vehicles affecting pedestrians and non-motor vehicles, while non-motor vehicle-pedestrian incidents represent a smaller proportion.

Integrating these findings with the eight problem categories identified earlier, the main causal factors are:

Subject Entity: Sudden rightward movement toward objects occurs to avoid obstacles on the main roadway, evade suddenly thrown objects, compensate for tracking distractions, overcome nighttime visual limitations, or facilitate proximity-based snatching. Retaliatory and loss-of-control incidents also predominantly involve rightward movement toward objects.

Object Entity: Slow reaction to rearward situations and vulnerability due to back-facing orientation.

To mitigate these issues, object entities must maintain facing orientation to enable defensive positioning and emergency response time. Literature [17] establishes that average human reaction times to frontal and rear stimuli are 0.15-0.5 seconds and 1-8 seconds respectively, representing several orders of magnitude difference in emergency handling capacity. Furthermore, mutual perception between subject and object entities can reduce incident severity. Based on these findings, we propose:

1) **Optimal pedestrian walking paths** follow the left-side sidewalk (or roadside) in the direction of travel. This positioning requires observation only of approaching motor and non-motor vehicles from the opposite right side, eliminating vulnerability from both directions, as incidents involving vehicles crossing the centerline to the left side occur with low probability. This configuration also effectively deters driving-based tracking and snatching, as such crimes require tailgating and rear-position execution, forcing perpetrators to risk violation and accident during approach, thereby increasing operational difficulty. Traffic Regulation Article 61 specifies that pedestrians should walk on sidewalks or, where none exist, along the roadside. Therefore, this safer walking pattern merely requires changing the current right-side walking habit without violating existing regulations.

2) **Enhanced driver perception** occurs when pedestrians walk on the left,

as motor and non-motor vehicle operators face forward-oriented pedestrians on their right. This orientation improves detection of pedestrian features (particularly eye reflections at night), sounds, physical reactions, and movement intentions, increasing driver alertness and enabling earlier preventive action.

3) **Reclassification of human-powered transport** is recommended. Pedestrians and human-powered vehicles (bicycles, tricycles, etc.) should be grouped into a human-powered category governed by pedestrian regulations, providing equivalent safety benefits. Rear-impact injuries from human-powered vehicles to pedestrians occur with low probability and severity, while rear-impact incidents from motor vehicles cause significantly more severe accidents than side impacts. This reclassification would also reduce pressure on main roadways.

Current traffic regulations mandate that motor and non-motor vehicles keep to the right side. Therefore, implementing these recommendations requires modification of existing traffic rules.

[1] ZHENG Changjiang, SHEN Jinxing. Data mining application in highway toll data[J]. Science paper Online, 2008, 3(10):174-176. (in Chinese)

[2] ARI W, WISNU J, HANIEF AW, et al. Traffic big data prediction and visualization using fast incremental model trees-drift detection(FIMT-DD)[J]. Knowledge-Based Systems, 2016(93):33-46.

[3] XIN Kejun, LÜ Bin, WANG Zhongyu. Big Data Technology-based Online Urban Transportation System Laboratory Environment Design[J]. Journal of Transport Information and Safety, 2014, 32(2):86-89

[4] PAN Xiaomin. Mining Road Traffic Flows Based On Association Rules by Using Apriori Algorithm [J]. JORNAL OF SHA NGHAI UNIVERSITY OF ENGINEERING SCIENCE, 2013, 27(3):283-288. (in Chinese)

[5] JIA Yi, HUANG Haofeng. Application of SQL Server in Traffic Accident Data Analysis[J]. Traffic Informatization, 2012, 11:131-134. (in Chinese)

[6] ZHANG Xiaobo. Research and implementation of mass data transferred based SQLServer Python[J]. Railway Computer Application | Railway Comput Appl, 2012, 21(2):55-57. (in Chinese)

[7] FRE' DE' RIC L, PATRICIA D. Speed behaviour as a choice between observing and exceeding the speed limit[J]. Transportation Research Part F: Traffic Psychology and Behaviour, 2005, 8(6):481-492.

[8] NAATANEN R, SUMMALA H. Road user behavior and traffic accident[J]. Oxford: North-Holland, 1976, 8(5):100-104.

[9] LI Shanhu. Evaluation method of aggressive driving behavior[D]. Xi' an: Chang' an University, 2011. (in Chinese)

[10] NEALE V L, DINGUS T A, KLAUER S G, et al. An overview of the 100-car naturalistic study and findings[R]. Washington, D.C., America: National Highway Traffic Safety Administration, 2005.

[11] WEI Hua, ZHANG Wei. Study on factors determining driving safety in China and the US[J]. China Safety Science Journal, 2005, 14(9):24-28. (in Chinese)

[12] FENG Zhongxiang, LIU Jing, LI Yangyang, et al. Aggressive driving behavior selection model and its influencing factors sensitivity analysis[J]. Chinese Journal of Highway, 2012, 35(5):210-214. (in Chinese)

[13] Brehmer B. Variable errors set a limit to adaptation[J]. Ergonomics, 1990, 33(10-11):1231-1239.

[14] ZHONG Zufeng, LIU Weiming. Based on the realization of the networking toll data to predict traffic flow[J]. China Management Informationization, 2009, 12(2):59-60. (in Chinese)

[15] DU Yong, LI Jun. Research on the method of data integration and design of integrated database of transportation data center[J]. Transportation Science & Technology, 2013(2):147-149. (in Chinese)

[16] HUANG Yuda, WANG Chaojie. An Intelligent Transportation Information Association Mining Scheme Based Improved Frequent Tree Model [J]. Computer Digital Engineering, 2015(2):211-213. (in Chinese)

[17] ZHAO Runshuan, PING Zhao. Correlation analysis of adults'waist hip ratio and human reaction speed[J]. Modern Medical Journal, 2014, 42(2):125-127. (in Chinese)

Note: Figure translations are in progress. See original paper for figures.