APPLICATION OF TEXT MINING IN ANALYSING ROAD CRASHES FOR ROAD ASSET MANAGEMENT

Richi Nayak[1], Noppadol Piyatrapoomi[2] and Justin Weligamage[2]

[1]Faculty of Science and Technology, Queensland University of Technology, Brisbane, QLD 4001, Australia

[2]Road Asset Management Branch, Queensland Government Department of Main Roads Brisbane, Queensland, Australia

Traffic safety is a major concern world-wide. It is in both the sociological and economic interests of society that attempts should be made to identify the major and multiple contributory factors to those road crashes. This paper presents a text mining based method to better understand the contextual relationships inherent in road crashes. By examining and analyzing the crash report data in Queensland from year 2004 and year 2005, this paper identifies and reports the major and multiple contributory factors to those crashes. The outcome of this study will support road asset management in reducing road crashes.

Key Words: Text Mining, Road Crashes, Data Analysis

1 INTRODUCTION

Traffic safety is a major concern in many states around Australia and including the state of Queensland. Since year 2000, there had been 296 fatalities on average per year in Queensland as recorded by the Office of Economic and Statistical Research [1]. Data obtained for analysis in this paper shows that during the years 2004 and 2005, there were over 20,000 traffic crash investigation reports recorded involving Queensland motorists. The annual economic cost of road crashes in Australia is enormous - conservatively estimated at $18 billion per annum - and the social impacts are devastating [2]. The cost to the community through these crashes is very high. They also have a devastating impact on the emergency services and a range of other groups. In addition, it is inevitable that the insurance companies will have to increase the cost of premium to cover their ongoing cost of insuring those motorists and their vehicles. It is therefore in both the sociological and economic interests of society that attempts are made to identify the major and multiple contributory factors to those crashes.

Statistical analysis of road crashes is not a new realm of research by any means. For many years, road safety engineers and researchers have attempted to deal with large volumes of information in order to gain an understanding of the economic and social impacts of car crashes. The hope is, that with this understanding, more efficient safety measures can be put into place to decrease the number of future road crashes [3]. Various data mining and statistical techniques have been used in the past in the domain. Researchers have attempted to investigate crash analysis through ordinary statistical tables and charting techniques [4, 5].The issue with these techniques is that they limit human involvement in the exploration and knowledge discovery tasks. Researchers have also attempted advanced data analysis methods of data mining that include clustering, neural networks and decision trees to reveal relationships between distractions and motor vehicle crashes. Major focuses of research on road crashes are the use of data mining to analyse freeway or highway accident frequency, the development of models to predict highway incident durations and the use of data mining in the classification of accident reports [6, 7, 8]. Other studies include the use of data mining and situation-awareness for improving road safety; a comparison of driving performance and behaviour in 4WDs versus sedans through data mining crash databases [5] and a study in the safety performance of intersections [9].

These studies revealed some interesting results, however, they are unable to properly analyse the cognitive aspects of the causes of the crashes. They often opt to leave out significant qualitative and textual information from data sets as it is difficult to create meaningful observations. The consequence of textual ignorance results in a limited analysis whereby less substantial conclusions are made. Text mining methods attempt to bridge this gap. Text Mining is discovery of new, previously unknown information, by automatically extracting it from different written (text) resources. Text mining methods are able to extract important concepts and emerging themes from the collection of text sources. Used in a practical situation, the possibilities for knowledge discovery through the use of text mining is immense. To our knowledge, there is limited or no reputable studies

that have utilised text mining in this data domain, however, earlier studies in the field indicate a real need for textual mining in order to better understand the contextual relationships of road crash data.

This paper presents a text mining based method to better understand the contextual relationships inherent in road crashes. By examining and analyzing the crash report data in Queensland from year 2004 and year 2005, this paper identifies and reports the major and multiple contributory factors to those crashes. Analysis is performed to identify links between common factors recorded in crash reports. Of key concern are the causes of crashes, rather than the consequences. The outcome of this study will support road asset management in reducing road crashes. With those findings on hand, we hope it can be useful for reviewing the limitations of existing road facilities as well as planning better public safety measurements. Most importantly, implementing and continuing a long term public education on road safety issues especially amongst the young generations and male gender which historically are involved in a high proportion of road crashes each year.

2 TEXT MINING METHOD

Text mining is discovery of new and previously unknown information automatically from different text resources using natural language and computation linguistics, machine learning and information science methods [10]. The key element is linking of the discovered information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation [10]. Text mining methods include the steps of processing the input text data, deriving rules and patterns within the newly processed data and finally the evaluation and interpretation of the output rules and patterns.

2.1 Objectives

The focus of this paper is to determine the most common causes of road crashes so that appropriate measures can be taken by road asset management in the future to prevent these accidents from occurring. The objective is to investigate the nature of crashes and roads that result in the crashes reported by Queensland traffic accident investigators within the period 2004 to 2005. Performing text mining on the crash description gives information about the causes of a crash that cannot necessarily be categorized into any particular field within a database. The crash reports, when pre-processed and grouped into clusters, can reveal insights that may have formally been unrecognisable. The identified unusual and hidden relationships may be useful to government, businesses (insurance organisations, motoring associations) and individuals in better road asset management. As compared with simply having a quantitative data set, textual information can enable conclusions to be drawn from the circumstances that caused the accident as opposed to simply looking at what the accidents were.

2.2 Dataset

The data used for this analysis, is collected from files containing information related to road crashes from the state of Queensland. The two files supplied are actual reports produced by traffic accident investigators within the period 2004 to 2005. They contain data for each reported road accident in this period according to 29 attributes including date, time, location and road conditions of the crash. More specifically they are: Atmospheric, Carriageway, Crash_Description, Crash_Nature, Crash_Area, Crash_Date, Crash_Day_of_Week, Crash_Distance, Crash_Divided_Road, Crash_Landmark, Crash_Number, Crash_Speed_Limit, Crash_Time, District, Horizontal_Allignment, Lighting, Owner_ID, Roadway_Feature, Road_Section, Road_Surface, Traffic_Control and Vertical_Allignment.

Preliminary review of the dataset helped us in determining the most interesting and important attributes to be used in our text mining analysis. However, for the purposes of text mining the crash description was of highest significance. This is a character attribute containing data with values up to 403 characters. The attribute “Street 2” with data of for example: “Warrego highway” had over 241 reported crashes which is also an interesting piece of information to look into. It is noted that there were as many as over 80% crashes occurred with clear conditions as shown in the attribute “Atmospheric”, and over 65% cases occurred during “daylight hours”. Also, a large number of reported crashes occurred on sealed and dry road surface as shown in the attribute “Road_Surface”, only small number of crashes occurred on wet surface. Another interesting attribute that came out of our observations was “Owner_ID”. It represents the gender of the person who was involved in each of these described crashes and it appears that almost all of them are MALE. The assumption is that this is due to the limitation of this particular crash report data. It may or may not be a true reflection of the correct distribution of gender involved in road crashes in a broader sense.

2.3 Process: Step1 - Pre-processing

Pre-processing of textual information is a time-consuming task but is essential in order to achieve results that are of value to the users of the information. An initial scan through the data set identified a number of potential problems that will need to be addressed before any text mining could take place. The main cause of this is mainly due to noise and various inconsistencies between the different records, which could possibly be due to different forensic experts writing the notes.

Punctuation: Punctuation was often omitted or used extraneously. No consistent information was apparent in the use of punctuation. Therefore, to simplify the text mining all punctuation was removed and replaced with spaces. Specifically, the following characters were replaced:

~`!@#$%^**&*()_+-={}[]|\;':'<>?/,.**

Broken Words: The previous step resulted in some words with gaps. Also there were many gaps (spaces) between words that will need to be removed in order to obtain any value from the words. This is because gaps between words are not actually "new" words, and during the text mining process they will have a low frequency, meaning that they are unlikely to be used. These gaps were consequently removed in order to obtain a more accurate result. Some examples are "trave lling", "ro ad".

Inconstancy due to the user of abbreviation and different cases: Another problem encountered with the data set is that there are many inconsistencies in different records. An example of this would be "unit 1", where variants including "u1" and "unit one" were used throughout the data set. This presents a problem in the context of road crash text mining because names of roads and highways could be abbreviated in a multitude of ways. In order to provide any meaningful recommendations, abbreviations also had to be standardised. It has been agreed that most value can be provided to the end user if the full word was used. Another example of inconsistency was using lowercase and uppercase to representing the same word. Consequently data was transferred to lowercase to prevent text mining tools from separating the same words that started with either upper or lower case. If this action was not taken, same words would not be grouped, therefore misleading the results and the integrity of the analysis. Converting all the text to lower case meant there were less combinations to code for transforming abbreviations to their full descriptions.

Spelling mistakes: Spelling mistakes was another common problem encountered during the pre-processing phase. They were removed by filtering through each record and correcting mistakes through the data set. If spelling mistakes are not corrected, text mining tools do not recognise the word, nor would it group same words. Some examples are "uint", "unti", "utni" user frequently to write the word "unit".

Common phrases: Finally as part of the formatting functions, common phrases that comprise of more than a word were combined into a single word to assist the text analysis. The data was processed for words which were in close proximity to each other to create common combinations. This was important as combinations such as green light and police station do not have the same meaning if they were not combined. For instance the car could have been green and it crashed into a light pole, instead of the car went through a green light and was hit in the middle of an intersection. Table 1 presents examples of the phrases replaced with concatenated words.

Table 1: Example replacements for common phrases

Original Text	Replace With
Traffic light	red-light
red light	stop-sign
turning lane	turning-lane
stop sign	stop-sign
green light	green-light
give way	give-way
lost control	lost-control
police station	police-station
parking bay	parking-bay
failed to stop	failed-to-stop
road side	road-side
bruce highway	bruce-highway
round about	roundabout
towing a trailer	towing-a-trailer

2.4 Process: Step 2 – Text Mining

The process of text mining includes converting unstructured text data into structured data, clustering the crash reports to identify links between common factors reported in crash reports, and viewing the concept links. We employ the Leximancer tool [13] based on the bayesian theory [14] to assess each word in the dataset to predict the concepts being discussed. It learns which word predicts which concept (or cluster) and forms concepts (or clusters) based on associated terms. It thus positions clusters based on the terms that they share with other clusters. It constructs a conceptual graphical map by measuring the frequency of occurrence of the main concepts and how often they occur close together within the text. A concept is treated as a cluster. Each *term* appearing in the text data is analyzed to form a *concept* to allow blackbox discovery of patterns that may not otherwise be known. Concepts that are similar are merged and edited. For example, the concept list included concepts such as turn, turning and turned; direction, north, south, east and west; light and lights; approached and approaching; road and street; lane and lanes. Each of these combinations of concepts relates to the same thing and one word is merely a stem of the other. As a result these similar concepts are merged into one and are renamed to reflect the true meaning of the concept, for example: Day, Time, Years, and Week are merged into the single concept 'Time'. Some concepts are removed which may not be pertinent to the crash senario being analysed, for example, preceded and occurred. Many concepts are then put together to form a *theme*.

Table 2: Example stop words excluded from the standard stop-word list

Word Excluded from the stop list	Rational
Bald	"Bald" tyres may be a cause of an accident.
Hit	Hit may imply a collision.
Look, looking	Look and Looking may be referring to where a driver was looking when the accident occurred.
Fast	Speed which may be the cause of an accident.
Indicate, Indicated	Whether a driver indicated left or right.
Right	Turning and merging right as opposed to left may have more of an impact in collisions. (i.e. turning across traffic).
Following	Car could be following too closely to another vehicle.
Two	Could refer to Unit two or number of vehicles involved in the accident.

3 ANALYSIS AND RESULTS

3.1 Dataset examination

Data distribution of the data set is displayed in figures 1 to 5 showing some significant correlations as well as disassociations between various attributes. The atmosphere is usually clear when the crash occurred indicating weather condition was not a big factor in this dataset. The distribution of crash time is mostly in the afternoon especially between 3pm to 5pm. This is during afternoon peak times when drivers are tired from working all day. The area with speed limit of 60km/h was where most crashes occurred, following by the area with speed limit of 100km/h. This is expected as the majority of roads have a speed limit of either 60km/h or 100km/h. No traffic control showing as most important contributory factor for the crash in this dataset. The three most significant crash natures were angle, hit fixed obstruction/ temporary object and rear-end.

Figure 6 shows that over a quarter of the accidents (28%) are classified as "rear-end" which is a high proportion of the data given there are 14 categories for this attribute. The data in Figure 7 displays the count of accidents grouped by the characteristic of the road where the accident occurred according to the "Roadway Feature" attribute. The proportion of accidents that occur at some form of intersections (i.e. Cross, interchange, multiple road, roundabout, T junction and Y junction) is 95% (excluding Not Applicable). This would indicate that there might not be enough controls in place around intersections to avoid an accident.

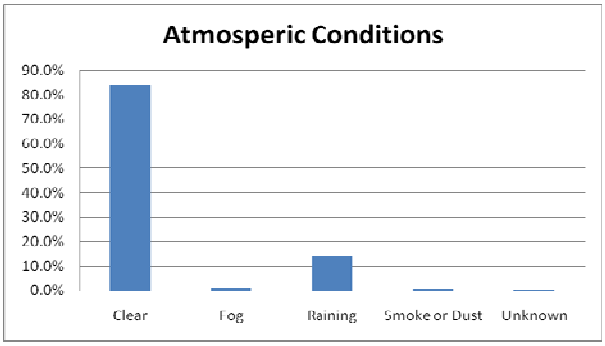

Figure 1 - ATMOSPHERIC attribute

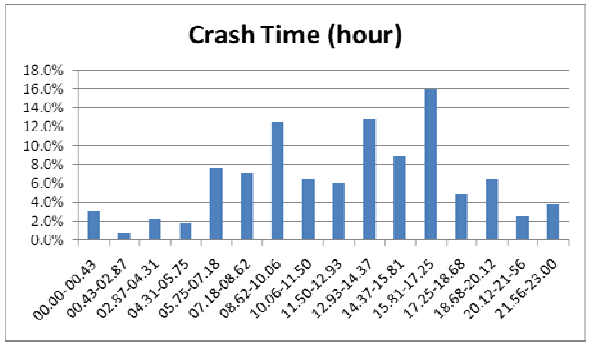

Figure 2 - CRAS_TIME attribute

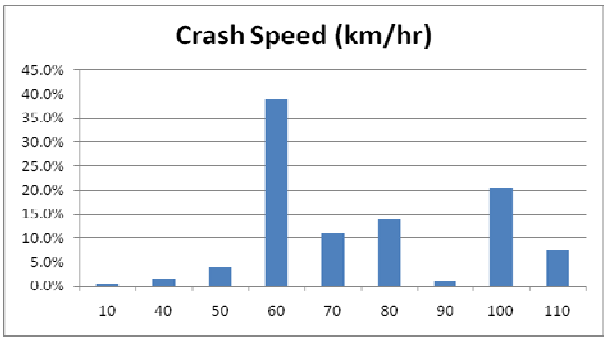

Figure 3 - CRAS_SPEED_LIMIT attribute

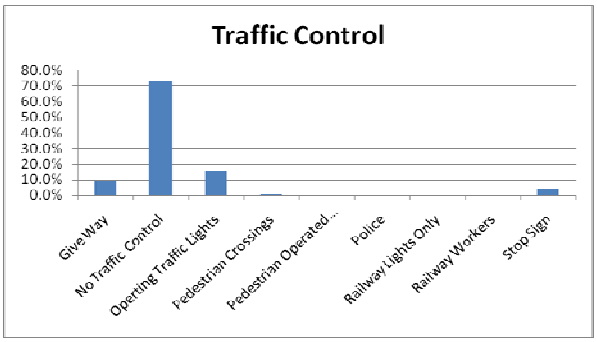

Figure 4 - TRAFFIC_CONTROL

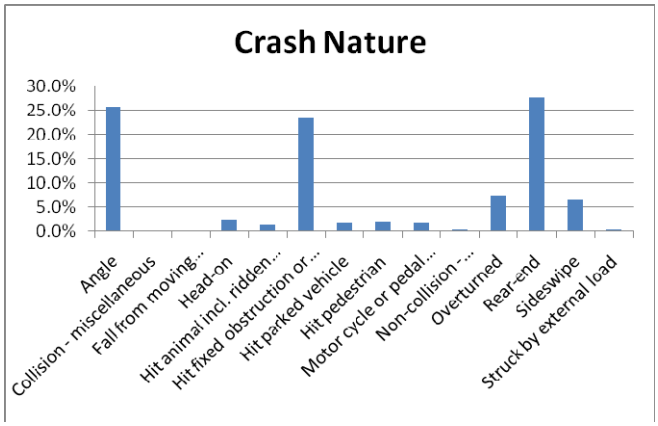

Figure 5 - CRASH_NATURE attribute

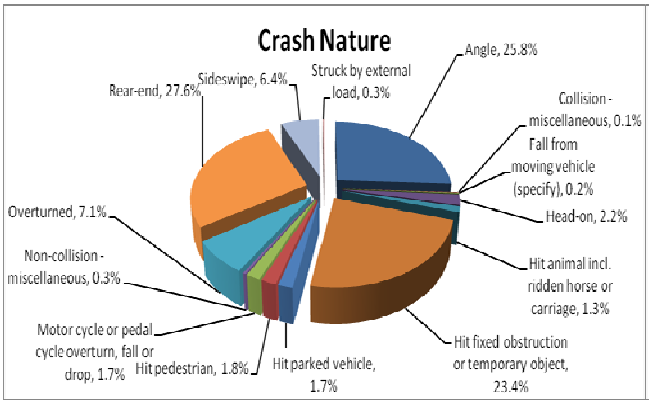

Figure 6 - Crash Nature Summary

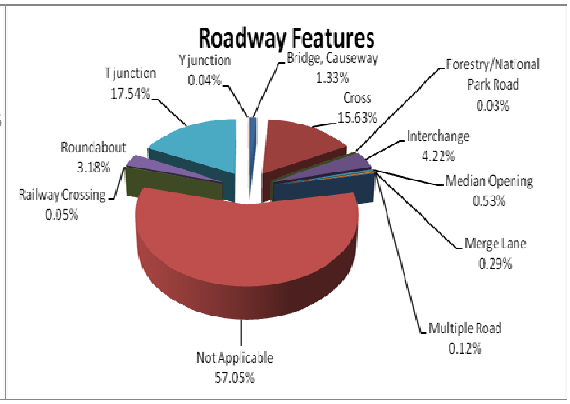

Figure 7: Roadway Feature Summary

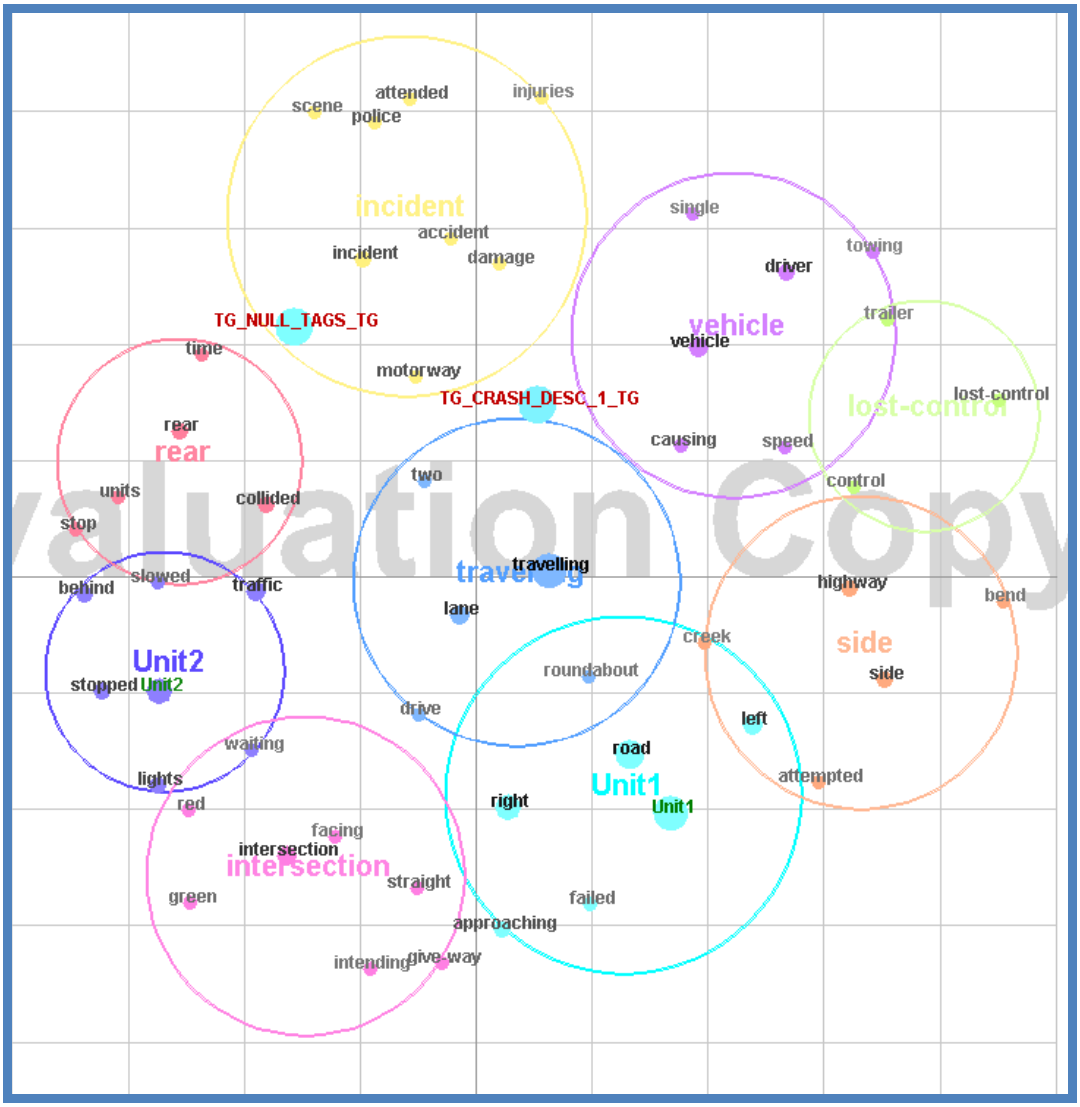

Figure 8 – The cluster map

3.2 Cluster Analysis

This Cluster Map in Figure 8 shows the different Clusters that were produced by the Leximancer text mining tool after the pre-processing had been preformed. Several clusters are immediately obvious by communicating possible causes for road accidents including intersections, rear-ending and loss of control. The list of concept terms generated by the Leximancer tool includes roundabout, intersection, traffic, bend, lane, injuries, left and right, give-way, rear and speed, among others. These terms alone give a good indication of possible causes of road accidents as they are the most frequently appearing terms in the sample text once stop words have been removed. The travelling, Unit 1, Unit 2, road, right, vehicle and intersection have the highest relative counts.

The two highest frequency concept terms are Unit 1 and Unit 2. Unit 1 occurs 12,774 times whilst Unit 2 only occurs 7,286 times. Similarly, the concept terms 'left' and 'right' appear 3111 and 6446 times respectively. An analysis into why 'right' might appear more than twice as many times as 'left' revealed that perhaps more accidents occur in right-hand lanes or whilst performing right-hand turns. Indeed, the relationship between 'right', 'left' and 'intersection' showed that the concept term 'intersection' will be accompanied by the concept term 'right' 73.5% of the time whilst it was only accompanied by 'left' 18% of the time.

An immediate assumption that could be made regarding the reason for Unit 1 appearing nearly twice as many times as Unit 2 would be that single vehicle accidents occur more frequently than multi-vehicle accidents. However, this assumption may not necessarily be true. For example, Unit 1 may just be repeated more times than Unit 2 within the same passage. Figure 9 indicates the strength of the relationship between *Unit 1* and all other concept words, whilst Figure 10 indicates the strength of

the relationship between *Unit 2* and all other concept words. The relationship from Unit 1 to Unit 2 is moderately strong whilst the relationship from Unit 2 to Unit 1 is significantly stronger. The relative count of the first relationship shows that the term 'Unit 2' (7286) is closely related to the term 'Unit 1' (7286) 100% of the time. However, the second relationship, shows that the term 'Unit 1' (12774) is only closely related to the term 'Unit 2' (7268) 57% of the time. This information indicates that 43% of the time a second vehicle is not involved. It is therefore possible to conclude that nearly half of all road crashes in this case study are single vehicle accidents.

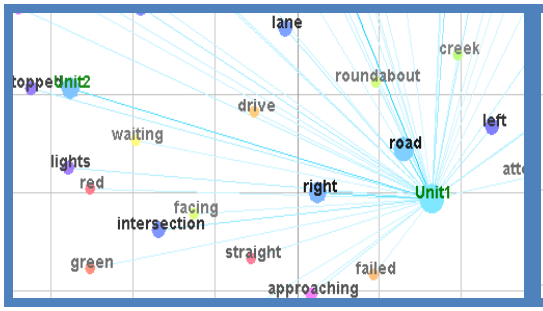

Figure 9: Unit 1 Concept

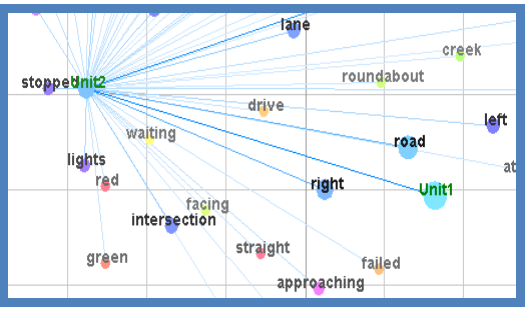

Figure 10: Unit 2 Concept

With the above discoveries in mind, the clusters can now be analysed to identify meaning in the grouping of concept words and their relative locations. The first meaningful clusters are the 'vehicle' and 'lost-control' clusters (as shown in Figure 11). These clusters appear in close proximity to each other, in fact overlapping, indicating a strong relationship between the two clusters. These two clusters include key words such as 'towing', 'trailer', 'speed' and 'lost-control'. One possible conclusion that could be drawn from these concept words is that drivers can often lose control of their vehicles when speeding. Another is that drivers can easily lose control of their vehicles when towing a trailer. The 'vehicle' cluster may also indicate a relationship between these conclusions and 'single driver' accidents or 'single vehicle' accidents.

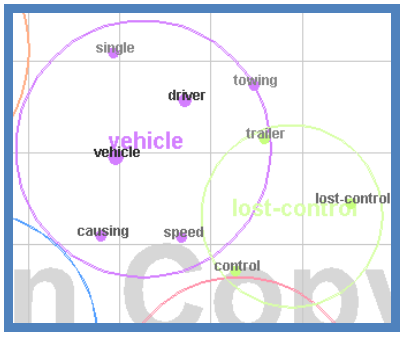

Figure 11: Driver Control Cluster

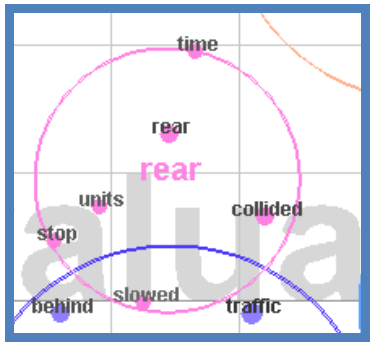

Figure 12: Rear-ending Cluster

The second meaningful cluster (as shown in Figure 12) is the 'rear' cluster which is also overlapping with the 'Unit 2' cluster. The concept words of the cluster include 'slowed', 'stop', 'time', 'collided' and 'rear'. These combination of words could indicate a scenario of rear-ending, a common form of car crash in suburban areas. The concept terms 'collided' and 'rear' alone would suggest this to be the case. However, this is also supported by the terms 'stop' and 'time' indicating that perhaps a vehicle could not 'stop in time' and as a result collided with the vehicle in front.

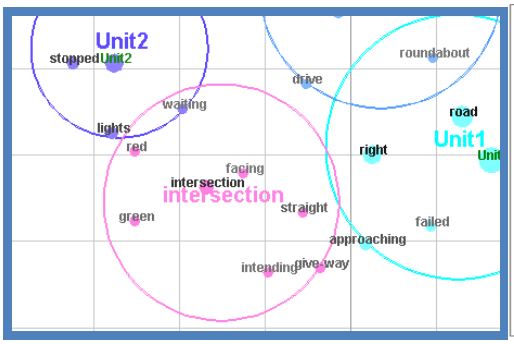

Figure 13: Intersection Cluster

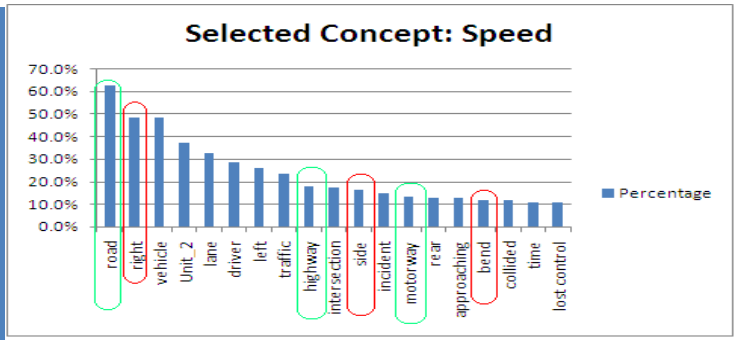

Figure 14: Speed Concept

The ‘intersection’ cluster (as shown in Figure 13) also seems to indicate quite meaningful information. An immediate observation that can be made is that this cluster overlaps with both the ‘Unit 1’ and ‘Unit 2’ clusters, indicating that perhaps accidents at intersections often involve two vehicles. The ‘intersection’ cluster includes interesting concept words such as ‘red light’, ‘green light’, ‘intersection’, ‘intending’ and ‘give way’. These key words might indicate that ‘giving way’ (or lack thereof) at intersections in perhaps a common cause of road accidents. This cluster may also suggest that traffic lights are often involved in crashes at intersections. Although this data alone does not tell us exactly how the traffic lights might be related to the accident one conclusion is that perhaps people are not stopping for red lights or simply do not see them. However, the relationship between traffic lights and causes of crashes is an area for further investigation.

Although these three clusters were perhaps the most meaningful, further conclusions could be drawn from the remaining clusters with further analysis. These three clusters in particular were chosen for analysis as they indicated possible *causes* for road accidents. The implications of these findings are discussed later.

One last area of interest is how road accidents are influenced by speeding. An analysis of the relationship between the concept word ‘speed’ and all other concept words indicates that perhaps speeding is a cause of more accidents in ‘low speed’ areas such as roads or streets rather than ‘high speed’ areas such as highways. As can be seen in Figure 14, the “speed” concept can be correlated with “road” or “street” (these two words were grouped in pre-processing) 63.6%. Whereas, highway and motorway only appear with speed 17.6% and 13.2% of the time respectively. Whilst the word ‘speed’ alone does not necessarily indicate speeding, it can be assumed that if the word were to appear in a crash report the speed must have been an influence factor.

4 DISCUSSION AND CONCLUSION

Several conclusions were drawn from the analysis conducted above. These conclusions involved: (1) the likelihood of a second vehicle being involved in an accident; (2) the likelihood of an accident when turning right as opposed to turning left; (3) the influence of towing a trailer in losing control of a vehicle; (4) the influence of speed in losing control of a vehicle; (5) a person’s inability to stop resulting in a rear-ending accident; (6) the likelihood of *more than one* vehicle being involved in an intersection accident; and (7) the influence of speed zone category in speeding accidents

From these conclusions, various recommendations can be made. Our proposed recommendations are as follows:

(a) greater awareness be raised regarding following another vehicle too closely or better known as tail-gating. Such awareness may help reduce the number of incidents related to rear-ending.

(b) determine new as well as improving existing controls to prevent these sort of rear ending accidents through signalling by the immobile vehicle. This would involve developing and improving methods of displaying that a vehicle is immobile to vehicles approaching it on either side. This includes for both the vehicle as well as for trailers that are attached to it to prevent the rear ending happening.

(c) further enhance future analysis of accident data, the improvement of information capture can be achieved by recording the presence/absence of right hand turning lanes at the intersection (for those accidents occurring at an intersection).

(d) determine new as well as improving existing roadway features. As mentioned in point (c), turning lanes are used to improve safety at intersections, however, if these are not able to be installed at certain intersections there may be a requirement to develop alternate controls. Another consideration is if turning lanes do not reduce accidents at intersections, which could be the subject of additional research.

(e) drivers purchasing trailers should be made aware of the difficult in controlling such vehicles and the implications associated with this.

(f) Speeding campaigns should target low speed zone areas rather than high speed zone areas and speed cameras should be utilized in low speed areas more often to discourage speeding in these problem areas.

(g) Furthermore, drivers should be reminded of ‘give way’ rules and there should perhaps be a greater focus on these rules during driving exams, particularly focusing on right-hand turns.

Table 4 lists recommendations for improvement in road assets according to the features that are highlighted during text mining analysis of the crash data set.

Finally, this paper has focused on the causes of road accidents and has not considered the consequences of such accidents but recognises that this is an equally significant area of concern. Whilst there is some information regarding injuries and damage to vehicles it is recommended that further research be conducted into the consequences of road accidents.

Table 4: Recommendation according to the features from the crash data set of interest

Other features	Recommendations for improvement (if possible)
Losing control of vehicle or rolling on embankment	Increased signage of accident-prone areas. Install road barriers if feasible.
Motorcyclist accidents	Increased regulations for gaining a motorcycle licence. Encourage motorcyclists to be more careful on roads at all times.
Speeding through intersections	This figure is concerning. More driver education is needed. Fixed speed cameras could be considered. Public awareness campaign might be required. Decrease speed limits if warranted.
Failure to obey signs.	Stricter consequences for violation of traffic regulations.
Collision with inanimate objects	Reflector strips on guardrails, other inanimate objects. Install parking lane for stopped vehicles.
Blood samples taken	-
Accidents at traffic lights	Increase visibility of traffic lights or install signage on approach to lights. Improve traffic lights at Southport and Nerang.
Police did not attend scene/minor accidents	-
Collisions due to right-hand turns	Consider installing a right-hand turn lane. Create signalised intersections.
Towing trailers	Educate drivers with long/ heavy loads how to manoeuvre vehicle properly.
Serious accidents requiring hospitalisation	Investigate each of these accidents separately.
Rear-end collisions	Remind drivers to keep a safe distance behind other vehicles at all times.

5 REFERENCES

1. Australian Government, Department of Infrastructure, Transport, Regional Government and Road Safety. (2008), Road Safety, http://www.infrastructure.gov.au/roads/safety/, Retrieved October/2008.
2. Queensland Fire and Rescue. (18/09/2002), Firefighters called to record number of road crashes, , http://www.fire.qld.gov.au/news/view.asp?id=207 , Retrieved October/2008
3. Abugessaisa, I. (2008). Knowledge discovery in road accidents database – Integration of visual and automatic data mining methods. *International Journal of Public Information Systems, 2008 (1),* 59-85. Retrieved October 20, 2008, from Emerald Insight database.
4. Gitelman, V. and Hakkert, A. S. (1997). The evaluation of road-rail crossing safety with limited accident statistics. *International journal of Accident Analysis Prevention, 29 (2),* 171-179. Retrieved October 20, 2008 from Emerald Insight database.
5. Gurubhagavatula, I., Nkwuo, J. E., Maislin, G., and Pack, A. I. (2008). Estimated cost of crashes in commercial drivers supports screening and treatment of obstructive sleep apnea. *International Journal of Accident Analysis & Prevention, 40 (1),* 104-115. Retrieved October 20, 2008 from Emerald Insight database.
6. Chatterjee, S. (1998). *A connectionist approach for classifying accident narratives*. Purdue University.
7. Li-Yen, C., & Wen-Chieh, C. (2005). Data mining of tree-based models to analyze freeway accident frequency. *Journal of Safety Research.*
8. Tseng, W., Nguyen, H., Liebowitz, J., & Agresti, W. (2005). Distractions and motor vehicle accidents: Data mining application on fatality analysis reporting system (FARS) data files. *Industrial management and data systems, 109 (9),* 1188-1205. Retrieved October 20, 2008, from Emerald Insight database.
   Journal of Safety Research.
9. Queensland University of Technology. (2008). Retrieved October 22, 2008, from QUT Centre for Accident Research and Road Safety: http://www.carrsq.qut.edu.au
10. Hearst, M. A. (1999): *Untangling Text Data Mining*. The 37th Annual Meeting of the Association for Computational Linguistics, Maryland, June 20-26, (invited paper). http://www.ischool.berkeley.edu/~hearst/text-mining.html. Accessed 20 April 2007.
11. Jain, A.K., M.N. Murty, and P.J. Flynn, *Data Clustering: A Review.* ACM Computing Surveys (CSUR), 1999. 31(3): p. 264-323.
12. Grossman, D. & Frieder, O. (2004): *Information Retrieval: Algorithms and Heuristics*. 2nd edn., Springer.

13. Smith A. E. Humphreys, M. S. 2006. Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping. http://www.leximancer.com/documents/B144.pdf Accessed 28 May 2007,

14. Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques.* San Diego, USA: Morgan Kaufmann

6 ACKNOWLEDGMENTS

We will like to thank CRC for Integrated Engineering Asset Management (CIEAM) to provide us the opportunity to conduct this case study. We will also like to thank students of ITB239 and ITN239: Enterprise Data Mining to conduct some of the experiments and Dan Emerson to assist us in reformatting the figures.